Federation of Agents: A Semantics-Aware Communication Fabric for Large-Scale Agentic AI

Lorenzo Giusti* **Ole Anton Werner** **Riccardo Taiello** **Matilde Carvalho Costa**

Emre Tosun **Andrea Protani** **Marc Molina** **Rodrigo Lopes de Almeida**

Paolo Cacace **Diogo Reis Santos** **Luigi Serio**

CERN, Geneva, Switzerland

Abstract

We present *Federation of Agents* (FoA), a distributed orchestration framework that transforms static multi-agent coordination into dynamic, capability-driven collaboration. FoA introduces *Versioned Capability Vectors* (VCVs): machine-readable profiles that enable the search for agent capabilities through semantic embeddings, allowing agents to advertise their capabilities, costs, and limitations. Our architecture combines three key innovations: (1) *semantic routing* that matches tasks to agents over sharded HNSW indices while enforcing operational constraints through cost-biased optimization, (2) *dynamic task decomposition* where compatible agents collaboratively break down complex tasks into DAGs of subtasks through consensus-based merging, and (3) *smart clustering* that groups agents working on similar subtasks into collaborative channels for k-round refinement before synthesis. Built on top of MQTT's publish-subscribe semantics for scalable message passing, FoA achieves sub-linear complexity through hierarchical capability matching and efficient index maintenance. Evaluation on HealthBench shows a 13-fold improvement over single-model baselines, with clustering-enhanced collaboration particularly effective for complex reasoning tasks that require multiple perspectives. The system scales horizontally while maintaining consistent performance, demonstrating that semantic orchestration with structured collaboration can unlock the collective intelligence of heterogeneous federations of AI agents.

1 Introduction

The landscape of artificial intelligence has evolved from single AI models to networks of specialized agents that plan, coordinate, and act over extended horizons [1, 2, 3]. This shift toward *agentic AI systems* represents a fundamental change in how we approach complex problem-solving with AI: rather than relying on a single model to handle all aspects of a task, we now orchestrate collections of specialized agents that can decompose tasks, maintain persistent context, and coordinate their efforts through structured communication towards a common goal. However, current agentic AI systems mainly rely on manually curated integrations and topic-based routing [4, 5], posing constraints on scalability as the heterogeneity of agents grows, and coordination complexity increases, limiting scalability and not addressing the fundamental operational question: *who can do what, at what cost, and under which policy constraints?*; preventing the realization of the "Internet of Agents" vision [6]. To address this, we introduce *Federation of Agents* (FoA), a semantics-aware communication fabric

*CERN, Corresponding to: lorenzo.giusti@cern.ch

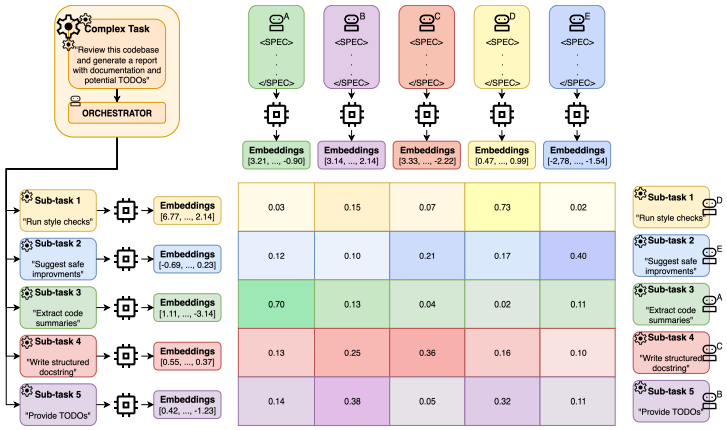

Figure 1: Orchestrator-driven Sub-task Decomposition and Semantic Routing.

that transforms agent coordination from static, topic-based routing to dynamic, capability-driven orchestration. At its core, FoA enables agents, tools, and data stores to advertise *Versioned Capability Vectors* (VCVs): machine-readable profiles that capture functional capabilities, performance characteristics, operational constraints, and security labels in a structured format.

Our Approach. FoA replaces static, topic-centric wiring with *dynamic, capability-driven orchestration*, aligning with calls for semantics-first coordination in agent ecosystems [7, 8, 4, 9]. Agents publish (VCVs), which are structured, versioned profiles embedded in a high-dimensional space, making capabilities searchable artifacts compatible with emerging interoperability efforts (e.g., Model Context Protocol (MCP)-based capability schemas) [10, 11]. We index VCVs using a sharded Hierarchical Navigable Small World (HNSW) index to support sublinear matching at scale while preserving nuanced distinctions among related skills [12]. At dispatch time, FoA applies *semantic routing* that couples profiles' similarities with policy checks and resource budgets (i.e., latency, bandwidth, energy consumption), rather than relying on keywords or static registries [13, 14]. Operational feasibility is ensured through transport-aware choices for IoT settings, where the Message Queuing Telemetry Transport (MQTT) protocol provides efficient and reliable delivery under constrained networks [15, 16, 17]. For *dynamic task decomposition*, FoA elicits candidate breakdowns from compatible agents and merges them into a consensual directed acyclic graph DAG, drawing on role-structured collaboration patterns from multi-agent systems [18, 19]. Finally, *intelligent orchestration* optimizes assignments over semantic fit and operational cost, supporting centralized modes for global objectives and decentralized modes for resilience at scale, consistent with observations from large agent populations and simulations [20, 21].

2 Federation of Agents Execution Flow

In this section, we analyze the life cycle of a single task handled by FoA. The framework orchestrates the end-to-end execution of a complex problem through a six-phase pipeline that captures decomposition, drafting, collaboration, and synthesis. Formally, given an incoming task t provided by the environment and a set of agents $\mathcal{A}$ equipped with VCV, the orchestrator A-0 establishes a DAG $G = (\mathcal{S}, E)$ of sub-task execution order whose vertices correspond to (sub-task, A-1) pairs and whose edges encode relational dependencies between sub-tasks. Each phase described below operates on G and updates its state until all nodes get detached from G by completing sub-tasks or by receiving a `DISPATCH` signal.

Sub-task decomposition, consensus, and assignment. Upon receiving a task t from the environment, Agent-0 embeds its natural-language description into the semantic space and queries the VCV index for candidate agents. Each compatible agent a_j returns a proposal consisting of a set of subtasks $\mathcal{S}_{a_j}$ and a set of dependencies E_{a_j} describing how those subtasks should be ordered. Agent-0 collects these proposals, merges them via a consensus mechanism, and validates acyclicity, thereby producing a global DAG $G = (\mathcal{S}, E)$ with $\mathcal{S} = \bigcup_j \mathcal{S}_{a_j}$ and $E = \bigcup_j E_{a_j}$. The orchestrator then solves the assignment problem introduced in Sec. 1: it computes scores α_{s_i,a_j} for each subtask-agent pair based on semantic alignment, policy compliance, resource fit, and specification similarity. Solving the resulting integer program yields an assignment matrix $\mathbf{X} \in \{0,1\}^{|\mathcal{S}|\times|\mathcal{A}|}$ that maps each subtask s_i to A-1 agents while respecting capacity constraints. FoA organises execution around G: leaves

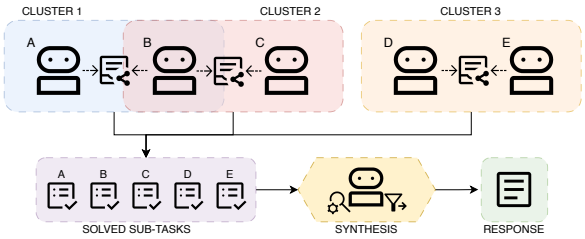

Figure 2: Collaborative refinement inside a high-similarity cluster.

(subtasks with no incoming edges) can begin immediately, whereas internal nodes s_i wait until all predecessors s_j with $(s_j \to s_i) \in E$ have reported completion. A `SYNTH` tool combines results from predecessors when triggering a downstream subtask, enabling partial results to propagate forward without blocking unrelated branches and facilitating fine-grained concurrency in large workflows.

First draft with resource access. Once assigned to a subtask s_i, each A-1 retrieves relevant context from its local resources, such as databases or external tools, via tool-use controllers embedded in A-1. Conditioning on this context and its specification embedding $\mathbf{e}_{a_j}$, the agent produces a first-draft answer $d_i^{(j)}$ for the subtask. This draft anchors subsequent refinement, which is posted to a cluster-specific MQTT channel associated with s_i. The retrieval step could implement an additional resource-aware step: the agent consults its resource vector $\mathbf{r}_{a_j}$ to adjust re-spawn parameters, ensuring that within clusters, A-1 size (in GB of GPU vRAM), throughput velocity (in tok/sec), context window of the channel, and maximum budget of available tokens to produce solutions do not exceed latency or energy constraints. As shown later in Sec. 3, we found it practical to set-up A-1 as a pre-aligned small language model ($\leq 20B$ params) for fast-feedback execution of multiple refinement rounds over long-term horizons [22, 23, 24].

Cluster formation via semantic similarity. Agent-0 groups the agents assigned to the same subtask into collaborative clusters based on their capability vectors and preliminary outputs. Concretely, for subtask s_i with assigned agents $\mathcal{A}_{s_i}$, we compute a similarity matrix combining (i) the cosine similarity of their capability embeddings $\mathbf{c}_{a_j}$, (ii) the cosine similarity of their draft embeddings and (iii) the overlap of their spec embeddings $\mathbf{e}_{a_j}$. Hierarchical clustering on this matrix yields clusters $C_1, \ldots, C_m$ of size chosen to balance diversity against coordination overhead. A dedicated cluster channel `foe/clusters/{cluster_id}/channel` is created on the MQTT broker for each cluster, allowing members to share messages without interfering with other topics. Fig. 2 illustrates the refinement process inside a high-similarity cluster.

Intra-cluster execution Within each cluster C_j, agents iteratively refine their drafts. At round r, every A-1 agent posts its current draft to the cluster channel and receives the drafts of its peers (Fig. 1). Agents critique and update their own answers by integrating insights from others, using simple majority voting or reputation-weighted aggregation to decide which components to adopt. Formally, let $\mathcal{M}_{s_i,C_j}^{(r)}$ denote the multiset of messages exchanged at round r; refinement continues for k rounds or until a consensus signal is triggered. Throughout this process, agents adhere to their Specs: they refuse to produce unsafe content, annotate uncertainties when appropriate, and propagate provenance metadata with each message. This collaborative refinement serves as a peer-review cycle designed to enhance factual accuracy and minimize inaccuracies.

Reporting to the orchestrator. When the agents in cluster C_j reach consensus on a refined answer $\hat{d}_i^{(j)}$ for subtask s_i, they emit a `TASK_COMPLETE` message on their cluster channel. This message contains the final answer along with evaluation metrics (e.g., confidence scores) computed by the cluster. Agent-0 subscribes to all cluster channels and listens for these completion signals. Upon receiving `TASK_COMPLETE` for s_i, it marks the node as finished in the DAG G and stores the result for use by downstream subtasks. If no consensus is reached within a predefined timeout, A-0 either reassigns the subtask to another agent or accepts the highest-scoring draft according to its evaluation function, thereby preventing deadlock.

Result synthesis. After all clusters have reported completion, A-0 traverses G in topological order. For each subtask s_i, it invokes the `SYNTH` operator to combine the results of its predecessor subtasks with the refined answer for s_i: $\mathbf{sol}_{s_i} = \texttt{SYNTH}\big(\{\mathbf{sol}_{s_j} : (s_j \to s_i) \in E\} \cup \{\hat{d}_i^{(j)}\}\big)$. This operator may concatenate texts (default solution), resolve conflicting assertions across cluster outputs (i.e., `rebase`), or summarise divergent perspectives into a unified answer (i.e., `merge`). We implement

Table 1: HealthBench Hard subset scores by axis. Individual example scores can be negative, but the average score is clipped to zero.

Axis	Best single agent	Uncoord. ensemble	Random assign.	FoA
Communication quality	0.56 ± 0.03	0.58 ± 0.02	0.51 ± 0.02	0.61 ± 0.03
Instruction following	0.39 ± 0.05	0.45 ± 0.02	0.39 ± 0.02	0.52 ± 0.04
Accuracy	0.15 ± 0.02	0.21 ± 0.02	0.10 ± 0.02	0.39 ± 0.02
Context awareness	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.01
Completeness	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.10 ± 0.03
Overall	0.01	0.02	0.00	**0.13**

`SYNTH` via meta-prompting [25] by steering the internal chain-of-thought of A-0 [26]. Once all leaves in the DAG are executed and their results propagated forward, the final answer for the original task t is obtained at the root of G. A-0 then publishes the final solution on `foa/result` MQTT topic, updates the reputations of participating agents based on the quality of their contributions, and archives their updated VCVs for future routing decisions.

3 Evaluation and Experimental Results

We evaluate the FoA framework on OpenAI's HealthBench Hard [27], a comprehensive benchmark for assessing language models in healthcare contexts. HealthBench Hard comprises 1,000 multi-turn conversations between models and users (both healthcare professionals and patients), with responses evaluated against physician-written rubrics spanning 48,562 unique criteria. The rubrics assign positive or negative points depending on whether a response satisfies desirable or undesirable criteria; scores range between -10 and 10 and are combined into a per-example score by a model-based grader that has been validated against physician judgments. The overall HealthBench score is obtained by averaging per-example scores and clipping the mean to the range $[0, 1]$. Unlike traditional multiple-choice medical benchmarks, HealthBench contains an open-ended nature of real healthcare interactions through conversation-specific evaluation across seven themes (emergency referrals, context seeking, global health, health data tasks, expertise-tailored communication, responding under uncertainty, and response depth) and five behavioural axes (accuracy, completeness, context awareness, communication quality, and instruction following).

Main Results. Tab. 1 summarises the performance of FoA and the baselines on HealthBench Hard. FoA achieves an overall score of 0.13, a *13x* relative improvement over the best single agent baseline (Medgemma [28]) and a *6.5x* improvement over the uncoordinated ensemble. Random assignment performs markedly worse, underscoring the importance of capability-aware routing. FoA consistently outperforms baselines across all seven themes. The collaborative cluster protocol particularly benefits high-stakes questions where multiple perspectives improve accuracy and context awareness.

4 Conclusion & Discussion

We presented Federation of Agents (FoA), a semantics-aware communication fabric that enables dynamic, capability-driven orchestration of large-scale multi-agent AI systems. Through machine-readable Versioned Capability Vectors (VCVs) and cost-aware semantic routing, FoA transforms static topic-based coordination into a scalable infrastructure for efficient reasoning over extended horizons, solving a coordination problem in agentic AI regarding *who can do what, at what cost, and with what reputation?* In conclusion, the Federation of Agents offers a theoretically grounded and practically scalable approach to multi-agent coordination, transforming capabilities and constraints into a searchable and auditable substrate. Our smart clustering innovations demonstrate that collaborative refinement can significantly improve solution quality while maintaining computational tractability. Addressing core challenges in semantic routing, distributed orchestration, intra-agent collaboration, and trust management establishes a foundation for the next generation of collaborative AI systems. The 10x performance improvements on HealthBench, compared to the best single agent, validate the practical benefits of capability-driven orchestration with smart clustering. Inspired by the principles of CAFEIN® [29, 30], we invite the research community to build upon these contributions as we collectively advance toward more capable, trustworthy, and socially beneficial agentic AI ecosystems.

References

[1] Ranjan Sapkota, Konstantinos I Roumeliotis, and Manoj Karkee. Ai agents vs. agentic ai: A conceptual taxonomy, applications and challenges. *arXiv preprint arXiv:2505.10468*, 2025.

[2] Johannes Schneider. Generative to agentic ai: Survey, conceptualization, and challenges. *arXiv preprint arXiv:2504.18875*, 2025.

[3] Mourad Gridach, Jay Nanavati, Khaldoun Zine El Abidine, Lenon Mendes, and Christina Mack. Agentic ai for scientific discovery: A survey of progress, challenges, and future directions. *arXiv preprint arXiv:2503.08979*, 2025.

[4] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36:38154–38180, 2023.

[5] Yanwei Yue, Guibin Zhang, Boyang Liu, Guancheng Wan, Kun Wang, Dawei Cheng, and Yiyan Qi. Masrouter: Learning to route llms for multi-agent systems. *arXiv preprint arXiv:2502.11133*, 2025.

[6] Yuhan Wang, Shuo Guo, Yang Pan, Zhi Su, Fuxiang Chen, Tom H. Luan, Peng Li, Jun Kang, and Dusit Niyato. Internet of agents: Fundamentals, applications, and challenges. *arXiv preprint arXiv:2505.07176*, 2025.

[7] Dezhang Kong, Shi Lin, Zhenhua Xu, Zhebo Wang, Minghao Li, Yufeng Li, Yilun Zhang, Hujin Peng, Zeyang Sha, Yuyuan Li, et al. A survey of llm-driven ai agent communication: Protocols, security risks, and defense countermeasures. *arXiv preprint arXiv:2506.19676*, 2025.

[8] Yingxuan Yang, Huacan Chai, Yuanyi Song, Siyuan Qi, Muning Wen, Ning Li, Junwei Liao, Haoyi Hu, Jianghao Lin, Gaowei Chang, et al. A survey of ai agent protocols. *arXiv preprint arXiv:2504.16736*, 2025.

[9] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First Conference on Language Modeling*, 2024.

[10] Mohammed Mehedi Hasan, Hao Li, Emad Fallahzadeh, Gopi Krishnan Rajbahadur, Bram Adams, and Ahmed E Hassan. Model context protocol (mcp) at first glance: Studying the security and maintainability of mcp servers. *arXiv preprint arXiv:2506.13538*, 2025.

[11] Xinyi Hou, Yanjie Zhao, Shenao Wang, and Haoyu Wang. Model context protocol (mcp): Landscape, security threats, and future research directions. *arXiv preprint arXiv:2503.23278*, 2025.

[12] Yury A. Malkov and Dmitry A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):824–836, 2018.

[13] Nima Seifi and Manish Chugh. Multi-llm routing strategies for generative ai applications on AWS. `https://aws.amazon.com/blogs/machine-learning/multi-llm-routing-strategies-for-generative-ai-applications-on-aws/`, April 2025. AWS Machine Learning Blog.

[14] Subash Neupane, Sudip Mittal, and Shahram Rahimi. Towards a hipaa compliant agentic ai system in healthcare. *arXiv preprint arXiv:2504.17669*, 2025.

[15] Andrew Banks, Ed Briggs, Ken Borgendale, and Rahul Gupta. Mqtt version 5.0, March 2019. OASIS Standard.

[16] EMQ Technologies. Harnessing LLM with MQTT: A comprehensive technical overview for ai/iot integration. Whitepaper, EMQ Technologies, 2024.

[17] Nouf Saeed Alotaibi, Hassan I Sayed Ahmed, Samah Osama M Kamel, and Ghada Farouk ElK-abbany. Secure enhancement for mqtt protocol using distributed machine learning framework. *Sensors*, 24(5):1638, 2024.

[18] Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, et al. Metagpt: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations (ICLR)*, 2024.

[19] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008, 2023.

[20] Yiming Xiong, Jian Wang, Bing Li, Yuhan Zhu, and Yuqi Zhao. Self-organizing agent network for llm-based workflow automation. *arXiv preprint arXiv:2508.13732*, 2025.

[21] Jing Piao, Yanyan Yan, Jiawei Zhang, Ning Li, Jin Yan, Xin Lan, Zhiheng Lu, Zhaolei Zheng, Jinyu Wang, Dong Zhou, and Chen Gao. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. *arXiv preprint arXiv:2502.08691*, 2025.

[22] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[23] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*, 2023.

[24] Peter Belcak, Greg Heinrich, Shizhe Diao, Yonggan Fu, Xin Dong, Saurav Muralidharan, Yingyan Celine Lin, and Pavlo Molchanov. Small language models are the future of agentic ai. *arXiv preprint arXiv:2506.02153*, 2025.

[25] Mirac Suzgun and Adam Tauman Kalai. Meta-prompting: Enhancing language models with task-agnostic scaffolding. *arXiv preprint arXiv:2401.12954*, 2024.

[26] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[27] Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. Healthbench: Evaluating large language models towards improved human health, 2025. OpenAI technical report.

[28] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

[29] Diogo Reis Santos, Albert Sund Aillet, Antonio Boiano, Usevalad Milasheuski, Lorenzo Giusti, Marco Di Gennaro, Sanaz Kianoush, Luca Barbieri, Monica Nicoli, Michele Carminati, et al. A federated learning platform as a service for advancing stroke management in european clinical centers. In *2024 IEEE International Conference on E-health Networking, Application & Services (HealthCom)*, pages 1–7. IEEE, 2024.

[30] CAFEIN®: CERN's federated ai platform. `https://cafein.web.cern.ch`.