Agents of Intelligence: Lowering the Barrier to AI for Domain Scientists

Ravi Shende[2], Tainã Coleman[1], Hena Ahmed[2], Ishayu Ghosh[2], Ismael Perez[2] and İlkay Altıntaş[2]
[1]*Loyola Marymount University*, Los Angeles, CA, USA
[2]*San Diego Supercomputer Center, University of California, San Diego*, La Jolla, CA, USA

***Abstract*—Artificial intelligence is reshaping how science is done, yet most domain scientists face a steep barrier to adopting it. The landscape of models, platforms, and data pipelines is fragmented, and navigating it rewards tacit expertise accumulated over years of practice. We present Agents of Intelligence (AoI): a modular, composable, human-in-the-loop framework of AI agents that sense workflow context, automate routine decisions, recommend models and compute resources, and learn from usage. AoI is grounded in interviews with eleven Schmidt AI postdoctoral researchers spanning eight scientific domains, and it integrates with the National Data Platform to support the full research lifecycle from data acquisition through model deployment, making AI-assisted science accessible across domains.**

I. The Expertise Gap

We interviewed eleven Schmidt AI postdoctoral researchers at UC San Diego working in medical imaging, neuroscience, molecular modeling, ecology, genomics, earth systems, and quantum chemistry. Despite the diversity of these domains, the same four barriers appeared in every conversation. First, most participants had no formal machine learning training; they are domain experts being asked to act as ML engineers. Second, compute resource selection is opaque, and researchers routinely cannot answer which cluster, which GPU, or how much memory a given job requires. Third, data preparation and metadata friction consume weeks, as cleaning, formatting, and documenting data is unglamorous work with no clear stopping point. Fourth, workflows are fragmented, and the tools for each stage of an analysis do not talk to each other. Together these describe a gap of accessibility: the computational expertise required to use AI in science is tacit, unevenly distributed, and rarely written down. This motivates our central question. Can AI agents encode the tacit computational expertise that domain scientists lack and place it at their fingertips?

II. The AoI Framework

AoI is a modular framework of agents that act as context-aware research companions. Agents can sense workflow state and user intent, automate routine steps to reduce cognitive load, recommend models and resources, and learn from feedback across domains and users. Three principles govern how they combine. Composability lets agents plug into a workflow as needed, at the domain science, application, or cyberinfrastructure layer. Context-awareness adapts an agent to the task, the data, and the model at hand. Human-in-the-loop keeps agents suggesting and scientists deciding. AoI can be accessed via MCP, CLI chat, or a UI. Integration with the National Data Platform resolves recommendations into executable NDP JupyterHub workspaces.

III. Agents on the Framework

BanditWare addresses hardware selection. Scientists submit jobs without knowing which hardware fits. The conventional remedy, training on historical traces, presumes data most groups do not have. BanditWare instead uses a contextual multi-armed bandit algorithm that balances exploration and exploitation to learn online how workflows perform on various hardware configurations, progressively improving its suggestions. This yields better runtimes and less wasted compute without manual tuning or a large historical corpus.

Model Recommendation and Fine-Tuning addresses the absence of formal ML training. Given a dataset and a task, a multi-agent system ranks the top five pre-trained models, retrieves the best fit from a model hub, and prepares it for fine-tuning in the user's workspace. Ranking aggregates three perspectives computed by separate agents, weighting relevance and citation evidence, community adoption on the model hub, and recency, so that no single ranking heuristic dominates. Optionally, a fine-tuning agent then retrieves the smallest model among the top five and adapts its final layer to the data, trading some accuracy for a model that is ready to use in minutes.

IV. Evaluation

We evaluate BanditWare on a matrix squaring dataset of runtimes measured across five hardware configurations, with matrix sizes from 5,000 to 12,500 and runtimes from one to thirty minutes. Accuracy measures how often the selected hardware was in fact the fastest of the five, and RMSE compares predicted against true runtime per hardware.

For model recommendation, every (LLM, dataset, run) triple placed at least one expert-validated model in the top five across the EuroSAT and IMDb datasets and across the LLMs evaluated.

To evaluate model recommendation of AoI versus prompting an LLM directly, we compare the two strategies against 3 datasets of various tasks and modalities shown in Table I.

For AoI, we use the same internal LLM as the one we were testing against: Qwen3. We gave the same prompts and data to the AI models. Finally, we implemented the suggested models with an 80/20 train/test split, tuning hyperparameters

TABLE I
DATASETS USED IN THE EXPERIMENTS.

Dataset	**Modality**	**Task**
Iris	Tabular	Classification
Air Quality (UCI)	Time Series	Regression
Garbage Types (UCI)	Image	Classification

with cross validation when applicable. Table II shows the experiment setup, while Table IV shows the results. Ultimately, accuracies were highest in AoI suggested models, while both models' secondary suggestion of LSTM performed the best compared to their primary suggestions.

TABLE II
MODELS SELECTED BY EACH RECOMMENDATION METHOD PER DATASET.

Dataset	**AOI Recommendation**	**LLM Recommendation**
Iris	TabPFN	SVM
Air Quality	ARIMA, LSTM	LightGBM, LSTM
Garbage Types	ResNet50	EfficientNet50

TABLE III
FINAL TEST RESULTS FOR EACH DATASET AND RECOMMENDER.

Dataset	Recommender	Model	Metric	Test Value
Iris	AoI	TabPFN	Accuracy	1.00
Iris	LLM	SVM	Accuracy	0.97
Garbage	AoI	ResNet50	Accuracy	0.85
Garbage	LLM	EfficientNet50	Accuracy	0.84
Air Quality	AoI	ARIMA	RMSE	0.70
Air Quality	Both	LSTM	RMSE	0.61
Air Quality	LLM	LightGBM	RMSE	0.65

V. ONGOING AND FUTURE WORK

Two agents are under active development: a sandbox for agent creation using `SKILL.md` based specifications, and a data harmonization agent that prepares datasets for the federated NDP catalog. Planned work extends AoI to digital twin orchestration and adds a dataset finder and a knowledge graph builder.