SPIO: Ensemble and Selective Strategies via LLM-Based Multi-Agent Planning in Automated Data Science

Wonduk Seo[1*], Juhyeon Lee[2*], Yanjun Shao[3], Qingshan Zhou[2], Seunghyun Lee[1], Yi Bu[2]

[1] Enhans [2] Peking University [3] Yale University

** denotes equal contribution.*

1. Background & Motivation

LLM-based agents are increasingly used in data science, but many still rely on **rigid**, **single-path** workflows.

Such fixed workflows limit **strategic exploration** and make it difficult to recover from **weak intermediate decisions**.

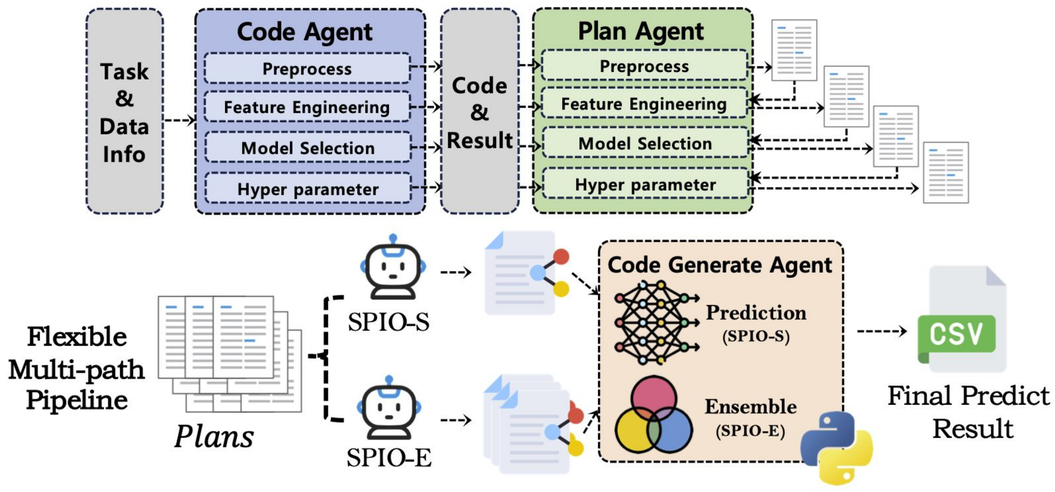
Automated data science requires **flexible planning** across key pipeline stages.

SPIO improves reliability by **exploring, evaluating, and integrating** diverse solution paths.

2. Dataset Details

Dataset	#Train	#Test	#Features	Label	Task	Null Ratio (Train/Test)
Kaggle Datasets						
Titanic	891	418	11	Survived	Classification	0.088 / 0.090
House Prices	1460	1459	80	SalePrice	Regression	0.067 / 0.068
Spaceship Titanic	8693	4277	13	Transported	Classification	0.021 / 0.020
Monsters	371	529	6	Type	Classification	0.000 / 0.000
Academic Success	76518	51012	37	Target	Classification	0.000 / 0.000
Bank Churn	165034	110023	13	Exited	Classification	0.000 / 0.000
Obesity Risk	20758	13840	17	NObeyesdad	Classification	0.000 / 0.000
Plate Defect	19219	12814	35	Class	Multi-class	0.000 / 0.000
BackPack	3694318	200000	9	Price	Regression	0.019 / 0.021
Rainfall	2190	730	11	Rainfall	Classification	0.000 / 0.000
Podcast	750000	250000	10	Listening_Time_minutes	Regression	0.031 / 0.031
OpenML Datasets						
Boston Housing	506	–	13	MEDV	Regression	0.000 / –
Diamonds	53940	–	9	Price	Regression	0.000 / –
KC1	2109	–	21	Defects	Classification	0.000 / –
SAT11	4440	–	117	Runtime	Regression	0.000 / –

3. Framework Overview: SPIO



4. Main Experimental Results

Methods	Titanic (ACC↑)	Spaceship (ACC↑)	Monsters (ACC↑)	Academic (ACC↑)	Obesity (ACC↑)	Kc1 (ACC↑)	Bank Churn (ROC↑)	Plate Defect (ROC↑)	House Price (RMSE↓)	Boston (RMSE↓)	Diamond (RMSE↓)	Sat11 (RMSE↓)
GPT-4o												
ZeroShot	0.7464	0.7706	0.7051	0.8269	0.8864	0.8499	0.8192	0.8640	0.1509	4.6387	540.3852	1447.1177
CoT (2022)	0.7440	0.7718	0.7108	0.8280	0.8859	0.8373	0.8719	0.8654	0.1447	4.6387	540.3556	1447.1177
Agent K V1.0 (2024)	0.7608	0.7810	0.6673	0.7879	0.8356	0.8490	0.8796	0.8209	0.1437	3.6797	415.8068	1338.4025
Auto Kaggle (2024)	0.7781	0.7753	0.7253	0.8275	0.8864	0.8422	0.8742	0.8351	0.1441	4.2510	417.9387	1365.7454
OpenHands (2024)	0.7528	0.7860	0.7189	0.8164	0.8815	0.8503	0.8830	0.8632	0.1379	3.4234	569.5314	1355.3216
Data Interpreter (2025)	0.7679	0.7870	0.7289	0.8297	0.8986	0.8482	0.8836	0.8828	0.1415	3.3937	420.9826	1376.4327
AIDE (2025)	0.7775	0.7949	0.7088	0.8340	0.8973	0.8567	0.8877	**0.8856**	0.1310	3.8445	444.0037	1319.8435
SPIO-S (Ours)	0.7847*	0.8010*	0.7316*	0.8341*	0.9071	0.8590	0.8877*	0.8836*	0.1310*	3.0312	404.6253	1290.9073
SPIO-E (Ours)	**0.7871***	**0.8034***	**0.7410***	**0.8359***	**0.9072**	**0.8687**	**0.8885***	0.8843*	**0.1298***	**2.9192**	**398.8893**	**1268.7817**
Claude 3.5 Haiku												
ZeroShot	0.7488	0.7904	0.7089	0.8281	0.8910	0.8594	0.7251	0.8657	0.1473	4.6387	543.1463	1446.9613
CoT (2022)	0.7488	0.7917	0.6994	0.8282	0.8904	0.8547	0.7454	0.8685	0.1467	4.6421	543.1463	1440.0644
Agent K V1.0 (2024)	0.7512	0.7928	0.7127	0.8288	0.8988	0.8639	0.8863	0.8813	0.1407	2.9710	546.9876	1445.1732
Auto Kaggle (2024)	0.7688	0.7884	0.7207	0.8182	0.8933	0.7427	0.8459	0.8243	0.1480	3.2183	789.4067	1288.3061
OpenHands (2024)	0.7727	0.7830	0.7051	0.8267	0.8894	0.8456	0.8656	0.8577	0.1445	3.1952	517.1757	1369.5845
Data Interpreter (2025)	0.7464	0.7940	0.6994	0.8282	0.8951	0.8578	0.8744	0.8642	0.1384	**2.8378**	565.1463	1350.7718
AIDE (2025)	0.7625	0.7893	**0.7316**	0.8335	0.8948	0.8706	**0.8887**	0.8856	0.1392	2.8556	560.4698	1273.8269
SPIO-S (Ours)	**0.7780***	0.7996*	0.7278*	0.8336*	0.9058*	0.8626	0.8812*	**0.8867***	0.1334*	2.8418	514.3608	1279.6327
SPIO-E (Ours)	0.7775*	**0.8027***	0.7297*	**0.8340***	**0.9066***	**0.8723**	0.8863*	0.8830*	**0.1332***	2.8409	**511.2089**	**1270.2695**
LLaMA3-8B												
ZeroShot	0.7410	0.7704	0.6880	0.8148	0.8793	0.8294	0.7783	0.8554	0.1521	4.6387	547.8507	1450.8444
CoT (2022)	0.7434	0.7819	0.7002	0.8181	0.8789	0.8279	0.7839	0.8570	0.1497	4.6387	546.4380	1425.4643
Agent K V1.0 (2024)	0.7512	0.7803	0.7013	0.8052	0.8855	0.8310	0.8671	0.8488	0.1458	4.0728	544.0589	1408.9078
Auto Kaggle (2024)	0.7415	0.7787	0.7018	0.8150	0.8775	0.8293	0.8505	0.8501	0.1469	4.1962	542.8794	1372.6835
Data Interpreter (2025)	0.7536	0.7830	0.7022	0.8187	0.8855	0.8362	0.8695	0.8613	0.1452	4.3823	538.0820	1378.5213
SPIO-S (Ours)	**0.7583***	0.7896*	0.7101	0.8208*	0.8935*	0.8424	0.8733*	0.8613*	0.1398*	3.6467	533.1050	1359.8238
SPIO-E (Ours)	0.7560*	**0.7907***	**0.7115***	**0.8238***	**0.8971***	**0.8474**	**0.8785***	**0.8739**	**0.1388**	**3.4386**	**524.1970**	**1353.8603**
Human Expert	-	0.8218	0.8072	0.8404	0.9116	-	0.9059	0.8898	-	-	-	-

Methods	RainFall (ACC) ↑	BackPack (RMSE) ↓	Podcast (RMSE) ↓
GPT-4o			
ZeroShot	0.8641	38.8250	12.3384
CoT	0.8655	38.8012	12.3158
Agent K V1.0	0.8691	38.7277	12.2849
Auto Kaggle	0.8879	38.7353	12.3042
OpenHands	0.8882	38.8392	12.4421
Data Interpreter	0.8884	38.7168	12.2549
AIDE	0.8932	38.8318	12.3371
SPIO-S (Ours)	0.8995	38.7071	12.1873
SPIO-E (Top 2 Ensemble)	**0.9004**	**38.6823**	**12.1072**
Claude 3.5 Haiku			
ZeroShot	0.8439	38.8138	12.3419
CoT	0.8497	38.8038	12.3242
Agent K V1.0	0.8526	38.7419	12.2750
Auto Kaggle	0.8719	38.7290	12.2388
OpenHands	0.8789	38.9418	12.4609
Data Interpreter	0.8731	38.7249	12.1948
AIDE	0.8812	38.9195	12.5523
SPIO-S (Ours)	0.8874	38.7071	12.1914
SPIO-E (Top 2 Ensemble)	**0.8923**	**38.6672**	**12.1823**
LLaMA3-8B			
ZeroShot	0.8420	38.8450	12.3700
CoT	0.8470	38.8320	12.3550
Agent K V1.0	0.8501	38.7890	12.2980
Auto Kaggle	0.8549	38.7840	12.3090
Data Interpreter	0.8564	38.7740	12.2760
SPIO-S (Ours)	0.8608	38.7520	12.2450
SPIO-E (Top 2 Ensemble)	**0.8738**	**38.7380**	**12.2278**
Human Expert	0.9065	38.6162	11.4483

- Evaluated on **12** benchmark datasets from **Kaggle** and **OpenML**.
- Tested with three LLM backbones: **GPT-4o**, **Claude 3.5 Haiku**, and **LLaMA3-8B**.
- Compared with strong baselines, including (1) ZeroShot, (2) CoT, (3) Agent K, (4) AutoKaggle, (5) OpenHands, (6) Data Interpreter, and (7) AIDE.

As a result …

- **SPIO-S** and **SPIO-E** consistently outperform baselines across classification and regression tasks.
- Achieves an average performance gain of **5.6%**, showing the effectiveness of sequential multi-path planning.

5. Ablation Studies

LLM	GPT-4o		Claude 3.5 Haiku		LLaMA3-8B	
	ACC/ROC ↑	RMSE ↓	ACC/ROC ↑	RMSE ↓	ACC/ROC ↑	RMSE ↓
Performance Comparison						
SPIO-S Top1	**0.8361**	**424.6737**	**0.8344**	**449.2421**	**0.8187**	**474.1788**
SPIO-S Top2	0.8308	448.2788	0.8339	469.6405	0.8025	480.1354
SPIO-S Top3	0.8248	501.7758	0.8203	504.4984	0.8052	496.9982
SPIO-S Top4	0.8193	454.9485	0.8176	596.6416	0.7931	498.5383
Ensemble Validation						
SPIO-S	0.8361	424.6737	0.8344	449.2421	0.8187	474.1788
SPIO-E Ensemble2	**0.8395**	**417.6800**	0.8365	**446.1131**	**0.8224**	**470.4087**
SPIO-E Ensemble3	0.8375	427.0149	**0.8476**	460.0835	0.8140	486.7630
SPIO-E Ensemble4	0.8341	429.5985	0.8323	524.2414	0.8014	486.3129
Module Impact Analysis						
(w/o) Preprocess	0.8310	434.4996	0.8276	472.4470	0.8331	497.3395
(w/o) Feature Eng.	0.8274	458.6899	0.8152	483.1569	0.8258	494.3421
(w/o) Model Select	0.8271	458.4007	0.8272	484.9608	0.8284	503.2921
(w/o) Hyper Param	0.8317	479.2854	0.8230	684.8432	0.8310	526.5389

6. Conclusion & Future Work

SPIO improves automated data science by exploring **multiple candidate strategies** instead of relying on a fixed single pipeline. Through sequential planning and integration across key stages, it achieves stronger and more robust performance on diverse benchmark datasets.

Future work will focus on making plan exploration more efficient, applying **SPIO** to more **complex** real-world scenarios, and improving **feedback mechanisms** for better pipeline selection and optimization.

Scan to read our paper!