

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

ACL 2026

Wonduk Seo^{1*}, Juhyeon Lee^{2*}, Yanjun Shao³, Qingshan Zhou², Seunghyun Lee¹, Yi Bu²

¹ Enhans ² Peking University ³ Yale University

** denotes equal contribution*

Speaker: Juhyeon Lee (PKU)

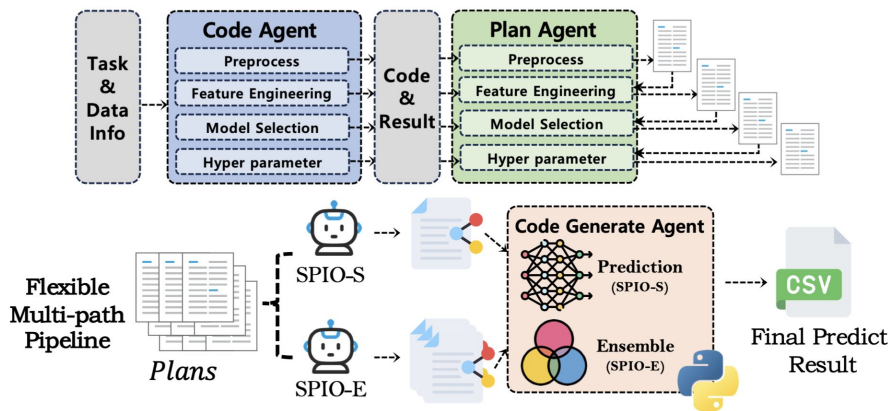
Table of Contents

1. Background and Research Gaps
2. Framework: **SPIO** (**S**equential **P**lan **I**ntegration and **O**ptimization)
3. Dataset Details
4. Experiments - Main Results & Ablation Studies
5. Additional Analysis - Qualitative Analysis & Case Study
6. Conclusion & Future Work

1. Background and Research Gaps

- Automated data science still relies heavily on **expert experience** and **repeated trial-and-error**.
- Existing AutoML methods are often limited by **predefined** search spaces, while current LLM-agent systems usually follow a **single workflow**.
- This makes it difficult to **flexibly explore** better data-processing and modeling strategies.

2. Framework Overview: SPIO



kaggle Titanic - Machine Learning from Disaster

Goal: It is your job to predict if a passenger survived the sinking of the Titanic or not. For each in the test set, you must predict a 0 or 1 value for the variable.

Metric: Accuracy (ACC)

Train/Test: 891 / 418 rows, 11 features

Label Column: Survived

ID	Gender	Age	Height	Weight	...
0	Male	24.4	1.69998	81.669950	...
1	Female	18	1.56	57	...
2	Female	18	1.71460	50.165754	...
3	Female	20.9	1.710790	131.274851	...
4	Male	31.6	1.914186	93.798055	...

GPT-4o Solution with Zero-Shot Prompting

- Load train/test CSV files.
- Split Target (label) from features in the train set.
- Preprocess features (median imputation + standardization).
- Train RandomForest on all training data.
- Predict on the test set and save the submission using the original ID.

ACC: 0.78

SPIO-S Solution (Best Single-path)

Load: Read train/test, set Survived, keep ID.

Preprocessing: Impute, encode, and interactions.

Features: Build Title/Cabin/Family features and predict missing Age.

Model: $R + \text{NB}$ soft-vote (1:1).

Hyperparameter: Optuna + stratified CV.

Output: Fit full train, predict test, save CSV.

ACC: 0.84

GPT-4o Solution with CoT Prompting

- Load train/test CSV files.
- Drop ID and label-encode Target in the train set.
- Split the training set into training/validation sets.
- Preprocess features (median imputation + standardization) inside a pipeline.
- Train RandomForest on the train split, then evaluate accuracy on the validation split.
- Predict on the test set and save the submission (uses row index as ID).

ACC: 0.80

SPIO-E Solution (Top-k Ensemble)

Load: Read train/test, set Survived, keep ID.

Preprocessing: Impute, encode, and interactions.

Features: Build Title/Cabin/Family features and predict missing Age.

Model: $R + \text{NB}$ soft-vote (1:1).

Hyperparameter: Optuna + stratified CV.

Output: Fit full train, predict test, save CSV.

ACC: 0.83

We propose **SPIO**, a multi-agent framework for automated data science.

SPIO generates multiple candidate strategies at each stage and sequentially integrates them into complete pipelines.

It supports both single best pipeline selection and ensemble-based pipeline optimization.

SPIO uses collaboration between the **Code Agent** and **Plan Agent**.

The **Code Agent** executes each data science step and reports results.

The **Plan Agent** generates diverse candidate strategies based on previous code, validation scores, and execution feedback.

Finally, **SPIO-S** selects the best pipeline, while **SPIO-E** ensembles multiple high-quality pipelines.

3. Dataset Details



Dataset	#Train	#Test	#Features	Label	Task	Null Ratio (Train/Test)
<i>Kaggle Datasets</i>						
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
<i>OpenML Datasets</i>						
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 / -

4. Experiments - Main 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	-	-	-	-

- **SPIO** improves performance by about **5.6%** over strong baselines on average.
- Compared with the top Kaggle leaderboard solutions, **SPIO** shows only about a **2%** performance gap.
- On the 2025 Rainfall dataset, **SPIO-E** reaches the **top 9%** of the leaderboard.

Dataset released after

2025

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

4. Experiments - Ablation Studies

We conduct ablation studies to analyze (1) the effect of plan ranking, (2) ensemble size, and (3) each SPIO module.

- **Performance Comparison** : Top-1 plans usually perform best
- **Ensemble Validation**: Ensemble-2 provides a strong balance between performance and stability
- **Module Impact Analysis** : Removing any module reduces overall performance.

LLM	GPT-4o		Claude 3.5 Haiku		LLaMA3-8B	
	ACC/ROC \uparrow	RMSE \downarrow	ACC/ROC \uparrow	RMSE \downarrow	ACC/ROC \uparrow	RMSE \downarrow
<i>Performance Comparison</i>						
SPIO-S Top1	0.8361	424.6737	0.8344	449.2421	0.8187	474.1788
SPIO-S Top2	<u>0.8308</u>	<u>448.2788</u>	<u>0.8339</u>	<u>469.6405</u>	0.8025	<u>480.1354</u>
SPIO-S Top3	0.8248	501.7758	0.8203	504.4984	<u>0.8052</u>	496.9982
SPIO-S Top4	0.8193	454.9485	0.8176	596.6416	0.7931	498.5383
<i>Ensemble Validation</i>						
SPIO-S	0.8361	<u>424.6737</u>	0.8344	<u>449.2421</u>	<u>0.8187</u>	474.1788
SPIO-E Ensemble2	0.8395	417.6800	<u>0.8365</u>	446.1131	0.8224	470.4087
SPIO-E Ensemble3	<u>0.8375</u>	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
<i>Module Impact Analysis</i>						
(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

5. Additional Analysis - Qualitative Analysis

- Evaluation objects

Participants:

Evaluation s

- Experiment 1

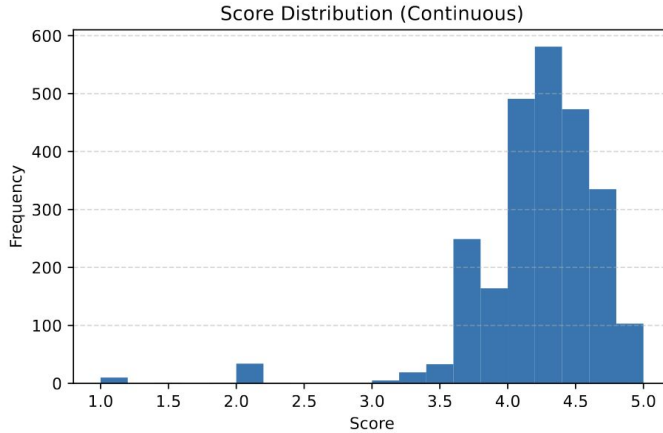
We evaluated

We also evalu

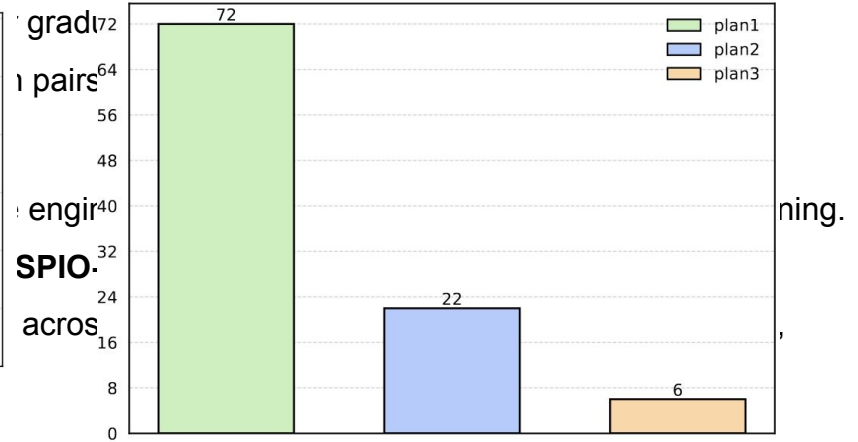
A 1–5 Likert

diversity & c

Experiment 1



Experiment 2



- Experiment 2: Forc

The Top-1 to Top-3

Participants were a

Dimension	SPIO-S	SPIO-E
Plausibility	4.38 ± 0.56	4.54 ± 0.29*
Interpretability	4.21 ± 0.45	4.37 ± 0.30*
Diversity & Coverage	4.10 ± 0.47	4.35 ± 0.30*
Usability & Trust	4.28 ± 0.46	4.55 ± 0.26*

y SPIO-E were anonymously presented to the participants.

and most suitable for each dataset.

5. Additional Analysis - Case Study

SPIO-5 | Input information description

@ Task Description

- Prediction Type
- Target Column
- Relevant Background

@ Raw data & Description Method

id	Gender	Age	Height	Weight
0	Male	26.414011	1.899289	84.899929
1	Female	18.000000	1.560000	57.000000
2	Female	18.000000	1.711460	50.165754
3	Female	20.952737	1.710730	131.274891
4	Male	31.641081	1.914186	93.798555
...
20753	Male	25.137087	1.766626	114.187096
20754	Male	18.000000	1.710000	50.000000
20755	Male	20.101026	1.819857	105.830481
20756	Male	28.852963	1.700100	83.501010
20757	Male	28.680376	1.819047	118.144888
...
20758	rows x 5 columns			

```
Data Description: {data_description}\nData Shape : {data_shape}\nData Column : {data_column}\nData length : {data_length}\nData Information : {df_info}\nDataframe head : {data_head}
```

Description of Train and Test set

Preprocess

@ Preprocess Code Agent

- Base preprocess code
- Preprocessed data description

@ Preprocess Plan Agent

- Method 1: Advanced Handling of Missing Values and Outliers
 1. Handling Missing Value
 2. Outlier Detection and Treatment
 3. Appropriate Circumstances & Reason for Recommendation
- Method 2: Encoding Categorical Variables and Feature Scaling
 1. Encoding Categorical Variables
 2. Feature Scaling
 3. Appropriate Circumstances & Reason for Recommendation
- Method 3: Feature Engineering and Dimensionality Reduction
 1. Feature Engineering
 2. Dimensionality Reduction
 3. Appropriate Circumstances & Reason for Recommendation

Feature Engineering

@ Feature Engineering Code Agent

- Base feature engineering code
- Preprocessed data description

@ Feature Engineering Plan Agent

- Method 1: Advanced Encoding of Categorical Variables
 1. Encoding Categorical Variables
 2. Appropriate Use Case & Reason for Recommendation
- Method 2: Interaction Features and Polynomial Features
 1. Interaction Features
 2. Polynomial Features
 3. Appropriate Use Case & Reason for Recommendation
- Method 3: Feature Selection and Dimensionality Reduction
 1. Feature Selection
 2. Dimensionality Reduction
 3. Appropriate Use Case & Reason for Recommendation

Model Selection

@ Model Selection Code Agent

- Base Model Selection code
- Validation Accuracy: 0.8972382787411689

@ Model Selection Plan Agent

Model 1: Random Forest Classifier

- Reason for Recommendation
- Appropriate Use Case

Model 2: Gradient Boosting Classifier

- Reason for Recommendation
- Appropriate Use Case

Model 3: Support Vector Machine (SVM)

- Reason for Recommendation
- Appropriate Use Case

Best Combination Method

@ Preprocess Plan:

- Method 2: Encoding Categorical Variables and Feature Scaling
- *Reason

@ Feature Engineering Plan:

- Method 1: Advanced Encoding of Categorical Variables
- *Reason

@ Machine Learning Model Recommendation:

- Model 1 – Random Forest Classifier
- Model 2 – Gradient Boosting Classifier
- *Reason

@ Hyperparameter Optimization Plan

- Method 3: Bayesian Optimization
- *Reason

Hyper Parameter Selection

@ Hyper Parameter Code Agent

- Base Hyper Parameter code
- Best Parameters:
{'max_depth': 20, 'min_samples_split': 5, 'n_estimators': 100}
- Best Score: 0.8982107825125398

@ Hyper Parameter Plan Agent

Grid Search with Cross-Validation (GridSearchCV)

- Reason for Recommendation
- Appropriate Circumstances

Random Search with Cross-Validation (RandomizedSearchCV)

- Reason for Recommendation
- Appropriate Circumstances

Bayesian Optimization

- (e.g., using the 'BayesianOptimization' package or 'Optuna')
- Reason for Recommendation
- Appropriate Circumstances

Final Code Generation

@ Import Library

- @ Load Data
- @ Separate features and target
- @ Preprocessing and Feature Engineering
- @ Define Models
- @ Hyperparameter tuning using Optuna
- @ Train final model with best hyperparameters
- @ Predict on test set
- @ Prepare submission

6. Conclusion & Future Work

SPIO improves automated data science by moving beyond a fixed single-path workflow.

It uses LLM-based multi-agent collaboration to generate, select, and integrate multiple candidate pipelines.

Experiments show that **SPIO** outperforms existing baselines on Kaggle and OpenML datasets, with **SPIO-S** selecting the best pipeline and **SPIO-E** improving stability through ensemble.

Limitations & Future Work

SPIO may not always find the globally optimal pipeline, and its results can be affected by the LLM, prompt design, and ranking quality.

Future work will focus on early pruning to reduce cost and extending **SPIO** to more real-world data science tasks.

Thank You!

Questions & Discussion are welcome



Scan to visit our ACL Paper