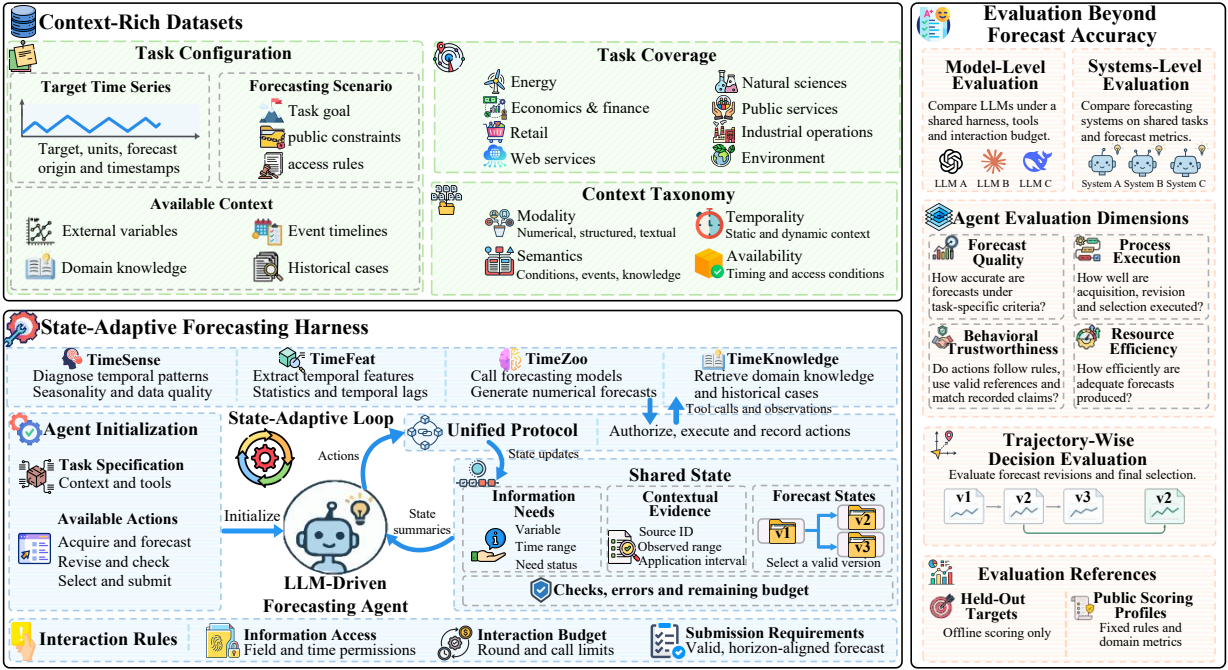

Context-Rich Datasets
Task Configuration
Target Time Series
Target, units, forecast origin and timestamps
Forecasting Scenario
Task goal
public constraints
access rules
Available Context
External variables
Event timelines
Domain knowledge
Historical cases
Task Coverage
Energy
Natural sciences
Economics & finance
Public services
Retail
Industrial operations
Web services
Environment
Context Taxonomy
Modality
Numerical, structured, textual
Temporality
Static and dynamic context
Semantics
Conditions, events, knowledge
Availability
Timing and access conditions
State-Adaptive Forecasting Harness
TimeSense
Diagnose temporal patterns
Seasonality and data quality
TimeFeat
Extract temporal features
Statistics and temporal lags
TimeZoo
Call forecasting models
Generate numerical forecasts
TimeKnowledge
Retrieve domain knowledge and historical cases
Tool calls and observations
Agent Initialization
Task Specification
Context and tools
Available Actions
Acquire and forecast
Revise and check
Select and submit
State-Adaptive Loop
Actions
Initialize
State summaries
LLM-Driven Forecasting Agent
Unified Protocol
Authorize, execute and record actions
State updates
Shared State
Information Needs
Variable
Time range
Need status
Contextual Evidence
Source ID
Observed range
Application interval
Forecast States
v1
v2
v3
Select a valid version
Checks, errors and remaining budget
Interaction Rules
Information Access
Field and time permissions
Interaction Budget
Round and call limits
Submission Requirements
Valid, horizon-aligned forecast
Evaluation Beyond Forecast Accuracy
Model-Level Evaluation
Compare LLMs under a shared harness, tools and interaction budget.
LLM A
LLM B
LLM C
Systems-Level Evaluation
Compare forecasting systems on shared tasks and forecast metrics.
System A
System B
System C
Agent Evaluation Dimensions
Forecast Quality
How accurate are forecasts under task-specific criteria?
Process Execution
How well are acquisition, revision and selection executed?
Behavioral Trustworthiness
Do actions follow rules, use valid references and match recorded claims?
Resource Efficiency
How efficiently are adequate forecasts produced?
Trajectory-Wise Decision Evaluation
Evaluate forecast revisions and final selection.
v1
v2
v3
v2
Evaluation References
Held-Out Targets
Offline scoring only
Public Scoring Profiles
Fixed rules and domain metrics