Supplementary Information

From public weather narratives to solar-market risk decisions using constrained language-model features

Yuqi Zhou[1] Xiangrui Meng[2,*] Jing Qiu[1]

Junhua Zhao[2]

[1]School of Electrical and Computer Engineering, The University of Sydney, Sydney, NSW 2006, Australia

[2]School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, Shenzhen 518172, China

[*]Correspondence: 222010046@link.cuhk.edu.cn

Supplementary Note: LLM weather-risk feature extraction

The LLM weather-risk feature cache was generated from National Weather Service Area Forecast Discussion (AFD) text for five forecast offices: `AFDLOX`, `AFDSGX`, `AFDHNX`, `AFDSTO`, and `AFDMTR`. For each local issue date and office, the feature builder used the synopsis, short-term, long-term, hazard-potential, and discussion sections when present, applied source-section filtering, and mapped the text into a bounded JSON-style schema.

The schema contains cloud severity, irradiance-reduction risk, storm risk, rain risk, fog or visibility risk, wind risk, heat risk, smoke or dust risk, morning risk, afternoon risk, evening risk, all-day risk, confidence, affected hours, rule-equivalent weather scores, and a short reason. All numeric fields are clipped to $[0, 1]$. The derived overall risk score is the maximum of the risk and time-of-day risk fields. The office-date cache was generated with the DeepSeek chat API using model `deepseek-v4-pro`, JSON-object output formatting, temperature 0.0, a 1600-token output limit, one API retry, disabled thinking output, up to three JSON parsing attempts, and deterministic post-processing. Office-level daily records are aggregated by date using maxima for risk fields, mean confidence, and summed product counts. The feature source is recorded as `deepseek_v4_pro_json_output`.

Quality-control item	Value
Forecast offices	5
Unique LLM-feature issue dates	1,462
Shifted feature dates used in the master table	1,462
Rows after required market/PV fields	25,854
2024–2025 test rows after lag filters	17,527
Selected-model training rows	8,303
PV capacity scale (MW)	21,637.75
Numeric risk/confidence completeness	100%
Numeric risk/confidence range conformance	100%

Supplementary Table S1: Experiment labels and abbreviations used in the manuscript.

Label or abbreviation	Meaning in this study
CAISO	California Independent System Operator.
SP15 or TH_SP15_GEN-APND	Southern California trading-hub price proxy used for day-ahead and real-time locational marginal prices.
PV	CAISO system-level photovoltaic solar generation.
DA and RT	Day-ahead and real-time market price series.
NWS AFD	National Weather Service Area Forecast Discussion text.
HRRR	High-Resolution Rapid Refresh numerical weather forecast fields.
NOAA Storm Events	County-level event labels used for stratified checks and event-day annotation outside the forecast feature sets.
MLP	Scikit-learn multilayer perceptron regressor used for PV forecasting.
Transformer	PyTorch sequence model used for real-time-price forecasting.
No-text	Forecast model using numerical weather, calendar, public forecast, market, and lag features but no NWS narrative-text features.
Rule-core or rule-text	Deterministic keyword-risk features mapped from NWS discussion text; the PV rule-core label denotes the cloud/rain-oriented rule feature set used as a matched deterministic text reference.
LLM cloud-rule	PV feature role that keeps the LLM-derived cloud-rule/irradiance-risk signal and excludes broader narrative fields less directly tied to solar irradiance.
LLM-rule	Price feature role that uses the rule-equivalent weather-risk scores returned by the constrained LLM extraction.
Scenario path	Residual-bootstrap PV and RT-price scenarios generated from a specified pair of forecast centers and training residual pools.
MLP no-text anchor	Deterministic MLP no-text PV forecast used as the common anchor in the main decision comparison.
LP-anchor hybrid	Final quantity $q_{\text{final}} = wq_{\text{LP}} + (1-w)q_{\text{anchor}}$, where q_{LP} is the stochastic LP quantity and q_{anchor} is the deterministic PV anchor.
Pure LP	The unanchored CVaR-regularized stochastic LP at $w = 1$.
Daily scenario CVaR penalty	CVaR term used inside the stochastic LP to regularize daily scenario losses when choosing q_{LP}.
CVaR95 loss	Reported ex-post hourly tail-loss metric: the negative average value over the worst 5% of realized hourly settlement-proxy values; lower values are preferred.
Imbalance proxy	Sum of absolute differences between the evaluated quantity and realized PV generation, reported in MWh or GWh.
Residual-bootstrap seed	Random seed controlling sampled training-residual rows used to generate PV and RT-price scenario paths.

Supplementary Note: forecast-seed results

The main text reports compact neural forecast comparisons. Supplementary Tables S2 and S3 report the corresponding five-seed summaries. Positive deltas in Supplementary Table S3 indicate lower error for the LLM-feature model.

Supplementary Table S2: Reported 2024–2025 forecast metrics across five initialization seeds. PV errors are MW and RT-price errors are USD/MWh.

Target	Model	Feature role	RMSE mean	RMSE SD	MAE mean	MAE SD
PV	MLP	No text	1186.03	12.58	633.22	11.73
PV	MLP	Rule-core text features	1221.09	13.48	668.62	13.32
PV	MLP	LLM cloud-rule feature	1167.35	5.17	614.89	4.09
RT price	Transformer	No text	18.27	0.49	10.90	0.47
RT price	Transformer	Rule-text weather scores	18.47	0.11	11.14	0.22
RT price	Transformer	LLM-rule weather scores	17.94	0.36	10.63	0.46

Supplementary Table S3: Paired five-seed forecast deltas for the main reported comparisons. Positive mean deltas indicate lower LLM-feature error than the reference.

Comparison	Metric	Mean delta	Mean improvement (%)	Paired p
PV LLM cloud-rule vs no-text	RMSE	18.69	1.57	0.0172
PV LLM cloud-rule vs no-text	MAE	18.33	2.88	0.0088
PV LLM cloud-rule vs rule-core	RMSE	53.74	4.39	0.00065
PV LLM cloud-rule vs rule-core	MAE	53.73	8.01	0.00036
RT LLM-rule vs no-text	RMSE	0.33	1.79	0.0518
RT LLM-rule vs no-text	MAE	0.27	2.42	0.0398
RT LLM-rule vs rule-text	RMSE	0.53	2.87	0.0499

Supplementary Note: forecast slice checks

The slice checks compare the LLM-feature model with its matched neural reference in the forecast instance used for downstream evaluation. Positive improvements indicate lower LLM-feature error.

Supplementary Table S4: Forecast slice metrics for the single cloud-rule downstream forecast instance used in the decision experiment. These values are not the five-seed means in the main forecast table. PV compares MLP LLM cloud-rule with MLP rule-core; RT price compares Transformer LLM-rule with Transformer rule-text.

Slice	Target	Hours	Reference RMSE	LLM RMSE	RMSE improvement
All test hours	PV	17,527	1236.60	1163.28	+5.93%
All test hours	RT price	17,527	18.54	18.09	+2.41%
Solar hours	PV	10,229	1616.54	1521.58	+5.87%
Solar hours	RT price	10,229	21.57	21.06	+2.41%
Extreme-event hours	PV	2,904	1338.72	1318.91	+1.48%
Extreme-event hours	RT price	2,904	25.53	25.13	+1.56%
Extreme solar hours	PV	1,694	1750.81	1726.29	+1.40%
Extreme solar hours	RT price	1,694	31.29	30.96	+1.04%
High LLM cloud-risk hours	PV	4,598	1301.25	1298.68	+0.20%
High LLM cloud-risk hours	RT price	4,598	21.13	20.73	+1.90%
High DA-RT spread hours	PV	4,382	1392.28	1385.03	+0.52%
High DA-RT spread hours	RT price	4,382	30.96	30.56	+1.30%

Supplementary Note: common-anchor settlement-proxy results

The downstream results use a prespecified validation selection over six LP-anchor weights: 0, 0.10, 0.25, 0.50, 0.75, and 1.00. The resulting strategies are frozen for the 2024–2025 test period. The main comparison fixes the deterministic MLP no-text PV anchor for both scenario paths so that no-text and LLM cloud-rule scenarios are compared under the same anchor convention. Stochastic downstream values use 10 matched residual-bootstrap seeds. The validation score in Supplementary Table S5 treats settlement-proxy value, hourly tail-loss exposure, and physical imbalance as co-equal validation dimensions. It averages six min–max-normalized components over the candidate weights: higher mean and minimum settlement-proxy value, lower mean and maximum ex-post hourly CVaR95 loss, and lower mean and maximum physical imbalance. This criterion selects an intermediate LP-anchor blend: larger LP weights increase mean validation value, but the associated increases in physical imbalance and upper-tail loss reduce the composite score. The score is computed on the LLM cloud-rule scenario path under the common MLP no-text anchor, and the selected $w = 0.25$ is then applied to both the LLM cloud-rule and no-text scenario paths in the matched test-period comparison.

Supplementary Table S5: Validation weight selection for the common no-text-anchor LLM cloud-rule path on the 1 October–31 December 2023 validation split. Higher validation score is preferred.

w	Mean value (M USD)	Min value (M USD)	Mean CVaR95 loss (k USD/h)	Max CVaR95 loss (k USD/h)	Mean imbalance (GWh)	Max imbalance (GWh)	Score	Rank
0.00	143.61	143.61	156.24	156.24	1704.5	1704.5	0.558	5
0.10	146.26	146.11	155.42	155.76	1714.3	1720.2	0.654	3
0.25	149.39	149.01	154.43	155.65	1746.2	1761.2	0.739	1
0.50	153.71	153.02	154.56	157.75	1817.0	1847.1	0.705	2
0.75	157.15	156.09	155.46	160.70	1905.5	1951.2	0.578	4
1.00	159.87	158.32	157.51	164.88	2008.5	2071.9	0.333	6

At $w = 0.25$, the validation score reaches 0.739, the highest value among the candidate weights. The next-highest candidate, $w = 0.50$, has a higher mean validation value but also larger maximum CVaR95 loss and physical imbalance, giving a lower composite score of 0.705. The selected weight therefore represents the validation-balanced point used for the aggregate test-period comparison and the event-day example.

Supplementary Table S6: Common-anchor SP15 settlement-proxy metrics. The anchor-only row is deterministic; stochastic hybrid and pure-LP rows report means across 10 residual-bootstrap seeds. CVaR95 loss is the ex-post hourly tail-loss metric defined in Methods. Pure LP denotes the unanchored CVaR-regularized stochastic LP at $w = 1$.

Strategy	PV scenario	RT scenario	Anchor	w	Value (M USD)	CVaR95 loss (k USD/h)	Imbalance (GWh)
MLP no-text deterministic anchor	–	–	MLP no-text	0.00	865.77	447.85	11251.0
No-text LP-anchor hybrid	MLP no-text	Transformer no-text	MLP no-text	0.25	869.12	459.94	11844.6
LLM cloud-rule LP-anchor hybrid	MLP LLM cloud-rule	Transformer LLM-rule	MLP no-text	0.25	881.41	453.43	11596.4
Pure no-text LP	MLP no-text	Transformer no-text	none	1.00	854.97	525.72	14110.3
Pure LLM cloud-rule LP	MLP LLM cloud-rule	Transformer LLM-rule	none	1.00	888.16	503.50	13436.3

Supplementary Table S7: Paired residual-bootstrap seed stability for the selected LLM cloud-rule LP-anchor hybrid relative to the no-text LP-anchor hybrid under the common deterministic MLP no-text PV anchor. Positive CVaR95 and imbalance rows indicate lower loss or lower imbalance for the LLM cloud-rule hybrid.

Reference	Metric	Mean	95% low	95% high	Paired p
Common no-text anchor	Value gain (M USD)	12.28	11.57	12.99	2.29×10^{-11}
Common no-text anchor	CVaR95 loss reduction (k USD/h)	6.51	6.21	6.81	3.38×10^{-12}
Common no-text anchor	Imbalance reduction (GWh)	248.2	234.5	261.9	1.55×10^{-11}

Supplementary Note: event-day mechanism example

The extreme-weather example in the main text traces the full forecast-to-decision pathway on an NOAA-labelled event day under the same common MLP no-text-anchor convention used in the aggregate decision comparison. Table S8 reports the Figure 4 event-day quantities for the selected $w = 0.25$ LLM cloud-rule LP-anchor hybrid relative to the deterministic MLP no-text anchor.

Supplementary Note: reproducibility summary

The pipeline writes a data audit, 2024–2025 test predictions, training residuals, forecast metrics, proxy-evaluation metrics, validation-selected hybrid summaries, paired seed summaries, forecast-slice tables, event-day example tables, and figure inputs. The release package stores paper-ready tables, figure inputs, run records, and the scripts used to regenerate the enrichment tables, figures, and table aggregations.

Supplementary Table S8: NOAA-labelled event-day mechanism example for the selected common-anchor LLM cloud-rule hybrid.

Date	Event labels	Reference	Value delta (k USD)	Imbalance reduction (MWh)	PV RMSE improvement (%)
2025-03-07	Heavy Rain	MLP no-text anchor	87.8	1948	10.3