Stout in Computer Science and Technology Studies (SCSTS)
ISSN: 2979-420X
DOI: https://doi.org/10.61424/d5mjap37
Journal Homepage: https://stoutjournals.org/index.php/SCSTS

Cost-Aware Tool-Routing Agents for Verifiable Agricultural Decision Support: An Offline Replay Benchmark

Kevin Zhang[1] ✉ **Yuxin Chen**[2] **and Daniel Wu**[3]
[1]*Computer Science, University of Southern California, CA, USA*
[2]*Statistics, University of California, Los Angeles, CA, USA*
[3]*Information Systems, New York University, NY, USA*
Corresponding Author: Kevin Zhang, **E-mail**: kevin.zhang0917@gmail.com

ARTICLE INFORMATION

Submitted: September 16, 2026
Accepted: October 13, 2026
Published: November 03, 2026
Volume: 2
Issue: 1

KEYWORDS

Agricultural decision support; tool routing; language models; evidence verification; cost-aware execution; offline replay; selective prediction.

ABSTRACT

Agricultural assistants must choose data and analytical tools without confusing source coverage, prediction, and verified evidence. We evaluated a constrained routing workflow using South Asian soil grids, Ethiopian cereal statistics, World Development Indicators, a flood archive, and the ND-GAIN composite index. The benchmark comprised 550 held-out requests: 420 supported queries, 70 unavailable-record probes, and 60 out-of-scope requests. Fixed execution, keyword rules, a supervised TF–IDF router, a selective variant, and a 77-million-parameter FLAN-T5-small router were compared under four operation budgets. At budget four, rules completed all supported tasks, TF–IDF completed 95.0%, and constrained FLAN-T5-small completed 38.6%. Rules used 1.15 source calls per request, compared with six for unrestricted fixed execution. Schema and identity checks rejected all 420 real-record substitution challenges, while removing verification produced incorrect accepted responses. A separate 284-observation forecasting test found that persistence achieved the lowest mean absolute error, 322.51 kg/ha. These results locate the main benefit in explicit evidence scope and efficient execution; they do not establish a general advantage for language-model routing. The evaluation concerns templated offline requests with resolved arguments, rather than autonomous field decisions.

1. Introduction

Agricultural decision support combines evidence with different spatial scales, time periods, and targets. Agricultural machine learning spans many such tasks (Liakos et al., 2018). Nepal-localized crop advisory and multimodal rice screening illustrate the need to retain the meaning of soil, climate, yield, and image inputs (Han & Zhou, 2023a, 2023b). Ethiopian crop classification and subnational portfolio ranking introduce further distinctions between predicting a crop label, ranking candidates, and estimating yield (A. Chen & Han, 2024; Han et al., 2026). An assistant must identify the requested operation before producing an answer.

Tool-using language models can connect these resources through a common interface. ReAct interleaves reasoning and external actions, while Toolformer learns API use (Schick et al., 2023; Yao et al., 2023). For agricultural applications, executable calls must also preserve location, year, property, units, and evidence coverage. A plausible number from the wrong source remains an incorrect response.

Cost must therefore be defined alongside answer quality. FrugalGPT and RouteLLM allocate model computation by query (L. Chen et al., 2024; Ong et al., 2024). Related systems optimize bitrate, rebuffering, and startup delay jointly (E. Wang & Wang, 2023), compare codec bandwidth and latency (H. Wang & Wu, 2024), or reduce beamforming complexity (H. Wang et al., 2020). These applications illustrate why resource measures should remain specific to the operation. Here, source-call budgets and measured routing time describe distinct costs.

Copyright: © 2026 The Author(s). This open access article is distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC- BY-NC-SA) License (https://creativecommons.org/licenses/by-nc-sa/4.0/), published by Stout Journals, London, United Kingdom

We evaluate fixed execution, rules, supervised text routing, and a compact instruction-tuned model over an offline agricultural tool catalogue. The questions are whether selective execution saves calls, whether evidence checks prevent incorrect acceptance, and when budget limits require abstention. The contribution is a controlled evaluation of these decisions; offline success does not establish farm-management benefit.

2. Literature Review

Agricultural advisory research increasingly combines prediction with explicit uncertainty and evidence presentation. Nepal crop-portfolio studies evaluate domain-shift-aware learning and conformal shortlists, while rice-disease triage considers clarification and abstention (Han & Zhou, 2025a, 2025b). Country-level yield and climate-resilience studies connect forecasts to grounded policy explanations and flood context (Han & Chen, 2024; H. Zhou & Han, 2026). These tasks motivate distinct tool contracts: a crop shortlist, disease-screening outcome, national forecast, and historical observation require different evidence and acceptance criteria.

General tool research provides complementary components. ToolLLM combines API retrieval and instruction tuning, and API-Bank evaluates executable planning and calls (Li et al., 2023; Qin et al., 2024). Retrieval-augmented generation supplies external information (Lewis et al., 2020); PAL delegates calculation to programs, while CRITIC incorporates tool feedback (Gao et al., 2023; Gou et al., 2024). Together, these approaches motivate separate evaluation of route choice, execution, and verification. A successful final response can conceal an inefficient route, while a correct route can still fail to produce a complete evidence bundle.

Transportation research makes the same separation concrete. Spatial retrieval connects crash hotspots with countermeasure evidence (Zhong et al., 2023), and graph forecasting supports corridor explanations and incident-aligned briefings (Ma & Zhang, 2026; Zhang & Zhang, 2024). Calibrated GraphRAG adds abstention to severe-crash screening and countermeasure ranking (Zhang & Ma, 2026). Digital-twin dispatching also evaluates sequential decisions with offline reinforcement learning (Zhang et al., 2025). These applications motivate preserving geographic identity and distinguishing descriptive evidence from action recommendations. Our context tool joins records without estimating treatment effects or learning a dispatch policy.

Prediction tools introduce another source of uncertainty. GPU-capacity forecasting studies calibrated uncertainty, and GPU-supply forecasting distinguishes readiness, scheduling delay, and capacity risk (He et al., 2023; Tu et al., 2024). Parcel-workload research examines unpaired visual and operational datasets (Ma et al., 2026), while enterprise credit assessment examines dynamic risk prediction (Wen, 2025). Across these settings, selecting a predictor and validating its forecast are different tasks. We therefore measure cereal-prediction error separately from whether the router retrieves the requested estimate.

Selective classification formalizes the coverage–risk trade-off, while calibration separates confidence from correctness (Geifman & El-Yaniv, 2017; Guo et al., 2017). Neither creates missing evidence. Validation must also respect temporal, spatial, and hierarchical dependence (Roberts et al., 2017), including leakage through preprocessing or repeated entities (Kapoor & Narayanan, 2023). Our evaluation holds out question phrasings and source groups, retains unavailable-record probes, and reports supported completion separately from refusal.

Evidence presentation is also an operational concern. Agricultural storytelling connects source material to visual interfaces (Liu & Han, 2025); traffic-sign explanation cards and advertising brief cards organize evidence and design decisions for inspection (Ma et al., 2025; Tu et al., 2025). Progressive PDF rendering and query-conditioned document streaming prioritize useful content under latency constraints (H. Wang et al., 2025; H. Wang & Wang, 2026). Distilled video-quality predictors provide compact representations for downstream use (B. Zhou et al., 2025). These designs suggest preserving source identity and uncertainty in concise outputs. The present benchmark evaluates structured evidence responses, leaving interface usability to a separate study.

3. Methodology

The four evidence families are summarized in Table 1. ACASA supplied 2,063 South Asian catalogue locations and 69 soil rasters; its methodology was released June 3, 2026, with dataset DOI 10.5281/zenodo.20037050. Sampling yielded 142,347 records, including 139,658 valid values. All 4,416 values at 64 Nepal locations were available. As with spatial soil products discussed by Poggio et al. (2021), gridded estimates retain resolution and uncertainty limits and are not new farm measurements.

Sampling used raster-specific transforms. The 66 layered rasters had 50,000-m cells in EPSG:7755; three other rasters used EPSG:4326 at different resolutions. Geographic catalogue coordinates were transformed before extraction, and missing pixels remained missing. Catalogue centres represented 2,006 distinct projected pixels. Evidence retained grid, property, depth, unit, source path, and pixel identity; no spatial soil predictor was fitted.

The Ethiopian table contained 1,820 records across ten regions, seven cereals, and 26 years during 1996–2022; 2003 was absent. There were 686 missing area entries, 716 zero yields, and 30 production-to-area yield discrepancies exceeding 1%. Source values and spellings were preserved. Supported retrieval excluded nonpositive yields and flagged discrepancies; returned statistics remained descriptive.

The WDI workbook contained 402,458 rows and columns through 2024. Indicator AG.YLD.CREL.KG supplied cereal yield in kilograms per harvested hectare. Country metadata excluded aggregates, leaving 217 economy identifiers and 10,264 observed values. The latest cereal observation was 2022, despite the later column labels.

The flood archive contained 5,130 records beginning January 1, 1985–October 6, 2021. Two repeated identifiers had nonidentical rows and were retained. Counts used primary-country labels and start years, without expanding secondary countries or merging label-specific totals. Replay excluded partial-year 2021 counts. Event coverage limits interpretation, as also emphasized by satellite flood assessments (Tellman et al., 2021). ND-GAIN supplied 5,236 values for 192 countries during 1995–2022 in a 2025 distribution. Its composite index was not decomposed into unavailable readiness scores (C. Chen et al., 2023).

Table 1. Evidence inventory and usable temporal scope

Source	Observed contents	Usable scope	Main constraint
D9 ACASA	2,063 locations; 69 soil rasters	Static; 2026 distribution	139,658 valid values; 64 Nepal locations
D3 Ethiopian cereals	1,820 region–crop–year rows	1996–2022, excluding 2003	716 zero yields; 30 arithmetic flags
D6 WDI cereals	10,264 nonmissing values	1961–2022 observations	217 economy identifiers; aggregate rows excluded
D7 flood archive	5,130 records	1985–2021 partial year	Replay counts restricted to 1985–2020
D7 ND-GAIN	5,236 values; 192 countries	1995–2022	Composite only; 2025 distribution

Soil metadata: dataset DOI 10.5281/zenodo.20037050; accompanying methodology DOI 10.12688/verixiv.3339.1. Observation years and distribution years describe different attributes.

Six executable tools supported eight plan labels (Table 2). Context combined national yield, flood count, and ND-GAIN for one country–year; stop invoked no source. Tools returned typed values, record keys, units, and provenance, with interval endpoints for forecasts. Timed replay used prepared in-memory indexes and cached predictions, excluding data preparation and model loading.

Table 2. Tool contracts and assigned execution costs

Plan or operation	Evidence returned	Cost units
Soil	Grid, attribute and depth value	1
Regional yield	Reported region–crop–year yield	1
National yield	WDI cereal yield by country and year	1
Flood	Count of primary-country archive records	1
GAIN	National composite index	1
Forecast	Frozen boosted-model estimate and interval	3
Context	National yield + flood + GAIN	3
Verify	Requested type and record identity check	1
Stop	Abstention	0

One cost unit is an assigned operation weight, not a monetary charge. Context uses three source calls. Forecast retrieval is charged three units even though its saved output is replayed.

Routers received question text only. Pre-resolved identifiers, years, properties, depths, and requested schemas went to execution and verification (Figure 1). The checker required one finite record of each requested type with a matching key and nonempty

unit/source fields. It did not independently recompute values or authenticate source truth. A separate scorer compared returned keys, types, units, sources, values, and interval endpoints with serialized reference answers, using absolute and relative tolerances of 10^{-9}.

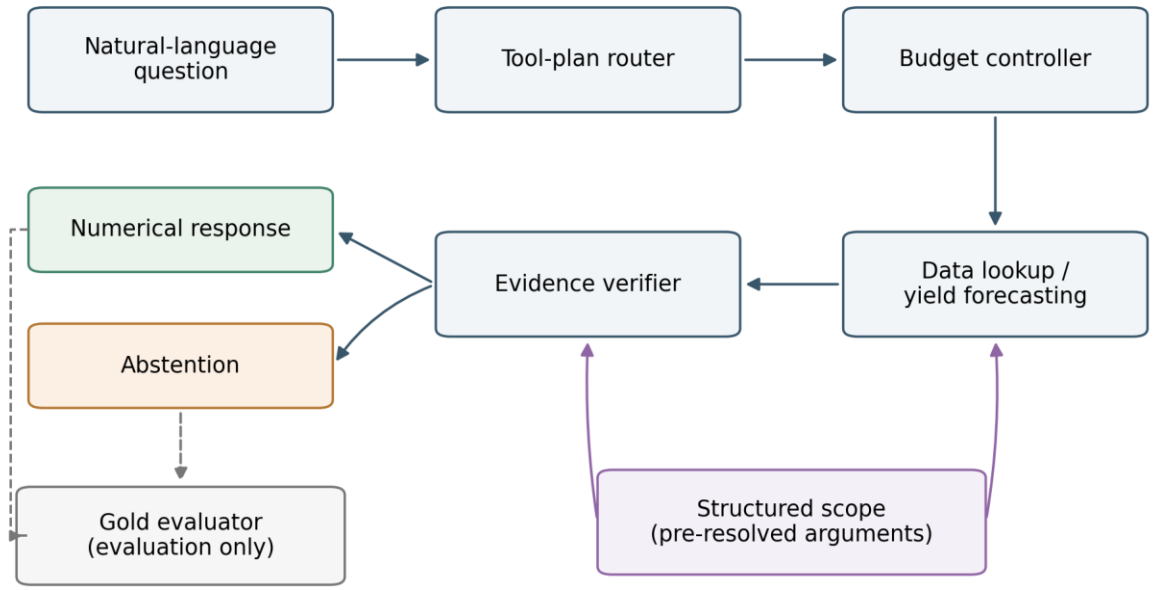


Figure 1. Routing and execution with a separate reference scorer. Pre-resolved scope is available to tools and verification; it is withheld from the text router. Acceptance certifies the implemented schema and identity checks.

Questions paired observed records with fixed English templates: three training, two development, and three test families per supported plan, plus six/four/six out-of-scope families. Seed 517 allocated approximately 50%/20%/30% of canonical ISO3 country groups to the respective splits, shared across country-level tasks. Ambiguous historical or composite flood labels were excluded from task selection. Soil groups were projected pixels; Ethiopian groups were region–crop pairs. Soil testing included 30 Nepal and 30 other locations. Table 3 reports task counts.

Table 3. Routing task allocation

Partition	Supported	Unavailable	Out of scope	Total
Training	1680	210	120	2010
Development	630	105	40	775
Test	420	70	60	550

The test partition has 60 supported and ten unavailable requests for each of seven evidence plans. Unavailable probes request 2025 records or soil depth 250–300 cm. No agricultural values are imputed for these probes.

Fixed all called soil, regional yield, national yield, flood, GAIN, and forecast in order; verification selected the requested subset. Rules followed fixed keyword patterns. TF–IDF used word unigrams/bigrams, sublinear term frequency, and logistic regression with C = 5, 1,000 maximum iterations, and seed 517. Selective TF–IDF returned stop below 0.65, the smallest development threshold with at most 1% accepted route error among candidates 0–0.90 in 0.05 steps. Selection did not establish calibrated probabilities.

The instruction-tuned baseline (Chung et al., 2024) was google/flan-t5-small, with 76,961,152 parameters, revision 0fc9ddf78a1e988dac52e2dac162b0ede4fd74ab. Its zero-shot prompt listed eight labels and descriptions before test exposure. Greedy inference used input/output limits of 512/12 tokens, FP32 CPU, batches of eight, and two threads. Invalid output triggered stop. After free decoding failed label formatting, an exploratory token-trie restriction allowed only the same eight labels; the semantic prompt was unchanged. Both runs were retained, without agricultural fine-tuning.

Budgets were 2, 4, 8, and 16 units. Guarded execution reserved one verification unit before each source call. Unaffordable calls stopped further execution; missing required evidence led to abstention. TF–IDF ablations removed verification or its reservation. Without reservation, raw evidence was emitted when checking became unaffordable. An identity challenge replaced returned evidence with different real records, using correct plans on 420 supported tasks. It tested record substitution, not arbitrary numerical corruption.

The forecasting experiment used three positive annual yield lags, their three-year mean, and the difference between the latest two lags. Exact year joins prevented gaps from becoming adjacent observations. Positive-target rows yielded 1,317 training observations in 2010–2018, 285 calibration observations in 2019–2020, and 284 test observations from 142 countries in 2021–2022. Countries recurred across periods; this was temporal evaluation.

Scikit-learn implemented three predictors (Pedregosa et al., 2011). Persistence returned lag one. Ridge predicted its residual using standardized features and regularization 10. Histogram boosting also predicted residuals: 150 iterations, learning rate 0.05, maximum 15 leaves, minimum leaf size 20, L2 penalty 1, seed 42. Outputs were clipped at zero. Training occurred once; 2022 predictions used observed 2021 yields, giving successive one-year-ahead forecasts with frozen parameters.

Intervals used absolute calibration residuals at order statistic ceiling[0.90(n + 1)], with lower endpoints clipped at zero (Angelopoulos & Bates, 2023). Coverage was evaluated empirically because temporal dependence and pooled countries limit exchangeability. Correctly retrieving a forecast was scored separately from predicting its observed target accurately.

Routing replay used 2019–2022 cached predictions: 27 of its 60 supported forecast queries concerned calibration-period outputs and 33 concerned test-period outputs. Calibration-period intervals were retrospective displays. Table 4 evaluates forecasting only on 2021–2022 observations.

Measures included supported completion, overall success, exact plan accuracy, answer coverage, evidence mismatch, incomplete answers, abstention recall, calls, cost, and time. Overall success included correct refusals on 130 unsupported queries. Mismatch meant an accepted item differed from requested reference evidence in type, key, unit, source, or value; partial but otherwise correct bundles were incomplete. Exact plan accuracy required the correct label, whereas completion allowed a sufficient superset. Paired differences used 1,000 bootstrap resamples of test entity groups, conditional on fixed templates. Batched routing time was divided by batch size and reported separately from replay.

4. Results and Discussion

Persistence had the lowest MAE, 322.51 kg/ha, and RMSE, 806.87 kg/ha (Table 4). Ridge and boosting increased MAE to 328.80 and 336.46 kg/ha. Boosting's narrower intervals achieved 90.14% coverage, versus 92.61% for persistence and ridge. Boosting remained the predefined learned-model tool, not a selected winner. These results reinforce the distinction between forecast quality and evidence retrieval underlying climate-yield explanation workflows (Han & Chen, 2024; H. Zhou & Han, 2026).

Table 4. One-year-ahead cereal-yield prediction on 284 test observations

Predictor	MAE	RMSE	Coverage %	Mean width
Persistence	322.51	806.87	92.6	1789.60
Ridge	328.80	849.34	92.6	1730.43
Boosting	336.46	807.84	90.1	1654.72

Errors and widths are kg/ha. Intervals target 90% coverage. Models share 1,317 training and 285 calibration observations; test years are 2021 and 2022.

At budget four, rules completed all supported tasks (Table 5). TF–IDF achieved 95.0% supported completion, 96.2% overall success, and 86.5% exact plan accuracy. Overrouting to context sometimes supplied a sufficient subset after checking. Constrained FLAN achieved 33.8% plan accuracy and 38.6% completion. Selective TF–IDF fell to 42.9% completion, showing poor transfer of its development threshold to held-out phrasings.

Free decoding produced invalid labels on all 550 test requests, so fallback abstention answered no supported tasks. Constrained decoding eliminated formatting errors but retained semantic routing errors. It is an exploratory format intervention rather than a prospectively selected superior policy.

Table 5. Routing and replay outcomes at budget four

Policy	Plan %	Support %	Success %	Error %	Calls	Cost
Fixed all	—	42.9	56.4	0.0	3.00	3.72
Rules	100.0	100.0	100.0	0.0	1.15	2.16
TF–IDF	86.5	95.0	96.2	0.0	1.31	2.23
Selective TF–IDF	49.5	42.9	56.4	0.0	0.48	0.84
FLAN constrained	33.8	38.6	53.1	0.0	1.00	1.71
FLAN free	0.0	0.0	23.6	—	0.00	0.00

Support denotes completion on 420 supported tasks; success includes 130 unsupported requests. Error is accepted evidence mismatch divided by all accepted responses. A dash denotes an undefined accepted-error rate when no answer is emitted. Calls count source operations only; cost includes verification.

Guarded policies rejected all 130 unsupported requests without accepting mismatched evidence. This reflects the supplied scope and trusted indexes: incorrect routes could still yield safe refusals. Figures 3–4 distinguish label confusions from family-level completion; they do not measure unrestricted language-model factuality.

Table 6. Supported-task completion across assigned budgets

Policy	Budget 2 %	Budget 4 %	Budget 8 %	Budget 16 %
Fixed all	14.3	42.9	85.7	100.0
Rules	71.4	100.0	100.0	100.0
TF–IDF	66.7	95.0	95.0	95.0
Selective TF–IDF	39.8	42.9	42.9	42.9
FLAN constrained	38.6	38.6	38.6	38.6

A verified lookup required two units, while context and forecast required four (Table 6; Figure 2). Rules completed 71.4% of supported tasks at budget two and 100% from budget four. Fixed all reached 14.3%, 42.9%, 85.7%, and 100% across increasing budgets; its curve depended on source order. At budget 16, fixed execution averaged six calls and 8.76 units, versus 1.15 calls and 2.16 units for rules: 75.3% fewer assigned units, without estimating monetary savings.

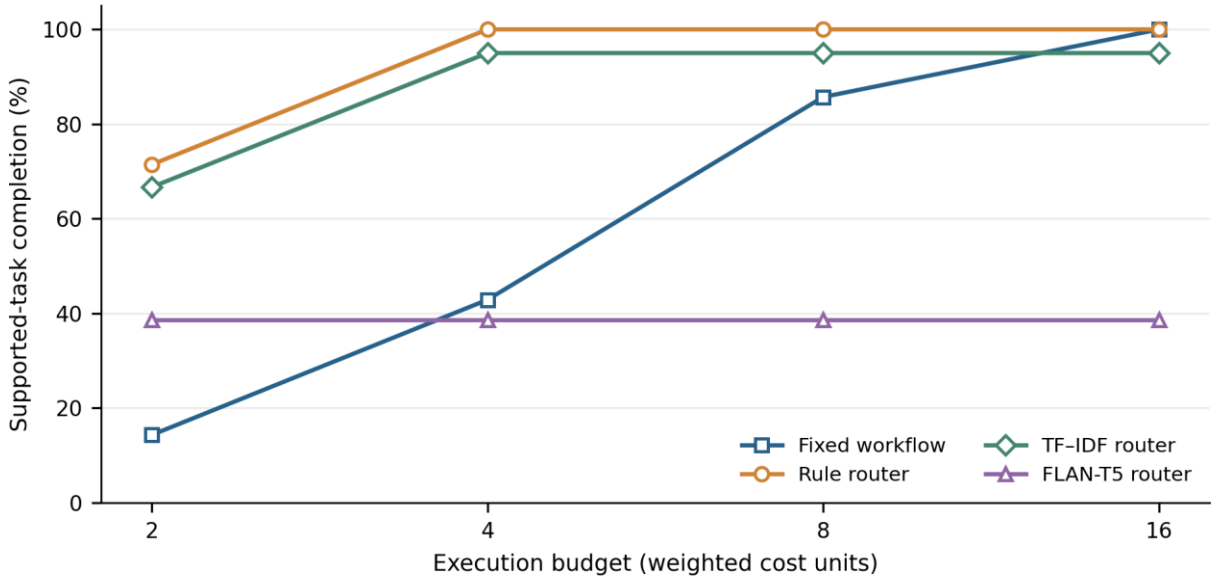


Figure 2. Supported completion versus execution budget. Cost weights are experimental settings; routing inference time is reported separately from these operation budgets.

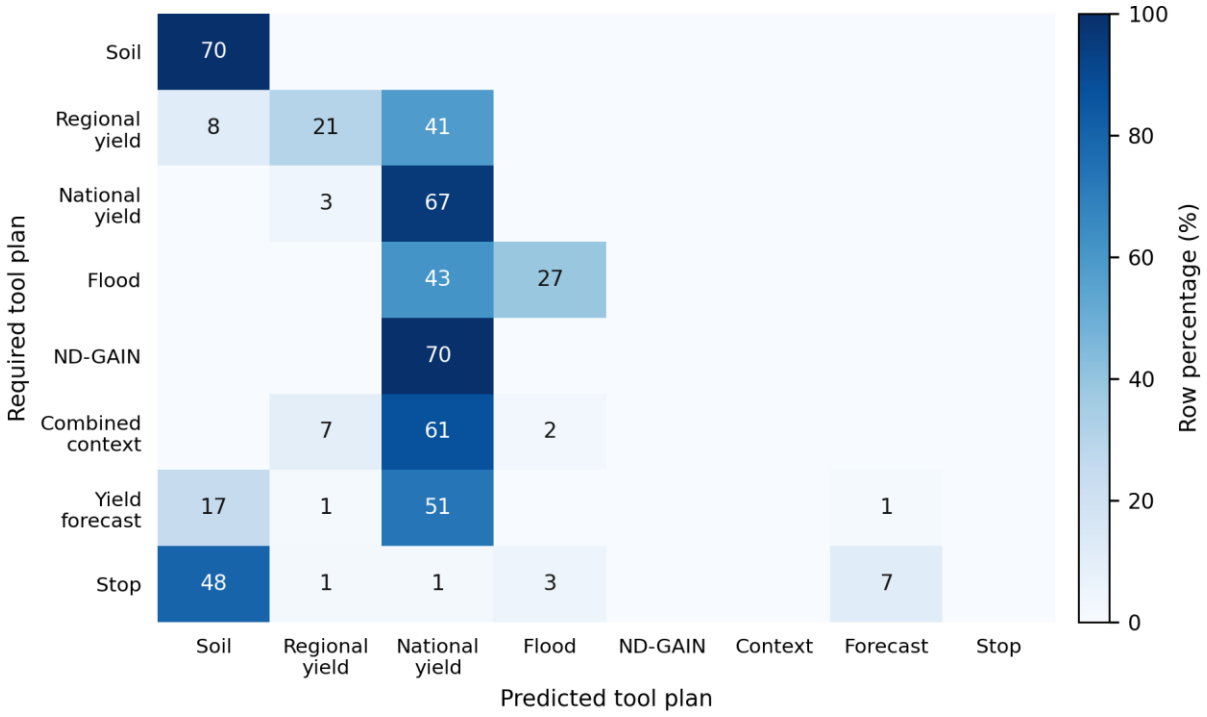


Figure 3. FLAN-T5-small plan confusion on all 550 test requests. Colors show within-row percentages; cell annotations are counts. Unavailable-record probes retain their source-plan label.

Table 7. Completion by supported evidence family at budget four

Family	Fixed %	Rules %	TF–IDF %	Selective %	FLAN %
Soil	100.0	100.0	100.0	100.0	100.0
Regional yield	100.0	100.0	100.0	61.7	35.0
National yield	100.0	100.0	100.0	31.7	100.0
Flood	0.0	100.0	100.0	25.0	35.0
GAIN	0.0	100.0	100.0	60.0	0.0
Forecast	0.0	100.0	65.0	0.0	0.0
Context	0.0	100.0	100.0	21.7	0.0

Each row contains 60 supported test requests. Missing-record probes are excluded from the completion denominator.

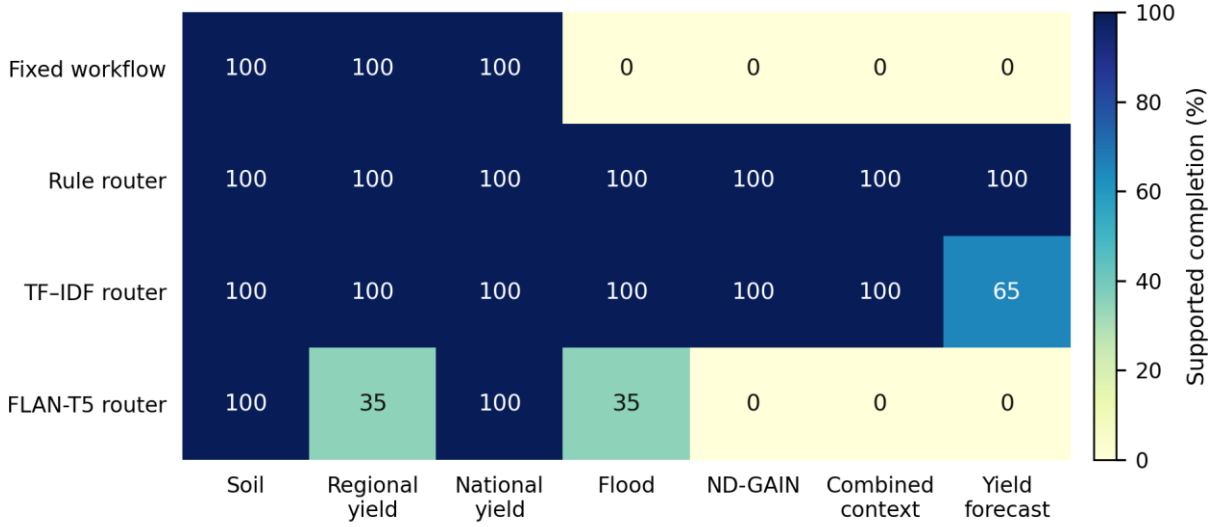


Figure 4. Completion by evidence family. Heterogeneity across the seven tasks exposes routing failures that an aggregate score can hide.

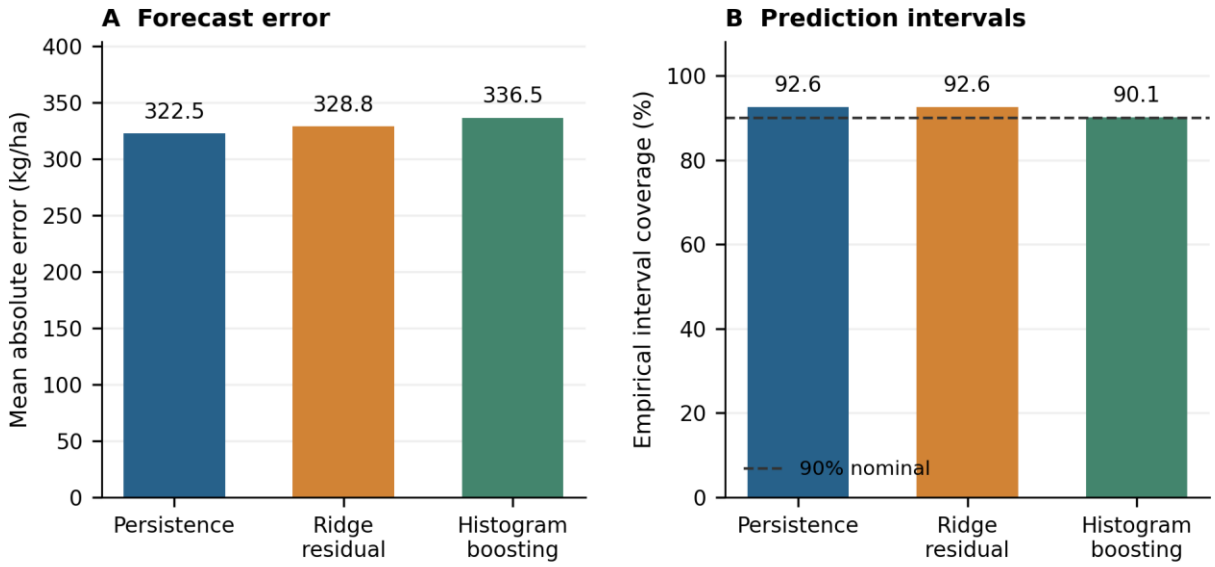


Figure 5. Predictor accuracy and empirical 90% interval coverage on the same 284 country–year observations. Persistence has the lowest MAE; boosting has the narrowest reported intervals.

At budget four, TF–IDF minus rules overall success was −3.82 percentage points, with a cluster-bootstrap 95% interval of −5.49 to −2.40. Rules benefited from a finite catalogue and recognizable vocabulary. This supports retaining a lexical baseline, without establishing its superiority for unrestricted agricultural dialogue. Family-level results and forecast quality are shown in Table 7 and Figures 4–5.

Removing verification produced 11.0% accepted mismatches at budget four (Table 8). Removing its reservation at budget two produced 5.8% mismatches and 16.7% incomplete responses among accepted responses. The checker rejected all 420 substituted bundles; unchecked execution accepted all 420 incorrectly (Figure 6). Correct identities carrying wrong finite values were outside this protection. Evidence contracts thus address a narrower property than the substantive safety of a recommendation.

Table 8. Verification and budget-reservation ablations

Condition	Budget	Support %	Accepted error %	Cost
Guarded TF–IDF	2	66.7	0.0	1.50
No reserve	2	66.4	5.8	1.53
No verification	4	88.8	11.0	1.47
Identity swap + check	16	0.0	—	—
Identity swap unchecked	16	0.0	100.0	—

The identity challenge uses 420 supported tasks and correct plans. Checked execution abstains on all substituted bundles, so accepted-error rate is undefined rather than zero.

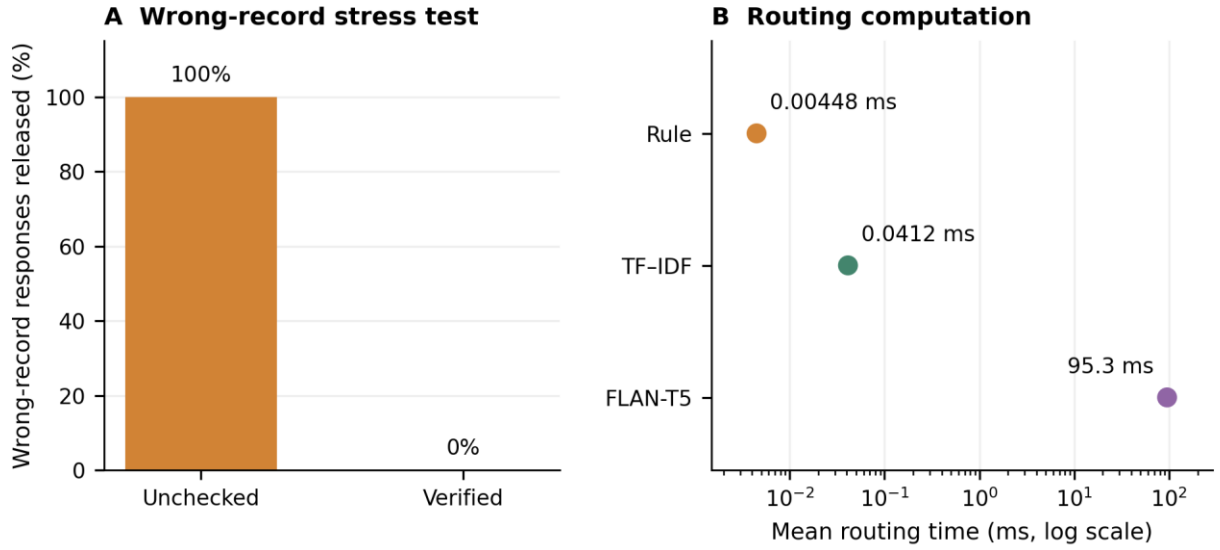


Figure 6. Real-record substitution rejection and router inference time. Identity challenges use correct plans; timing reports text routing separately from indexed tool execution.

Table 9 separates source-call budgets from routing computation. Prepared indexes and cached forecasts made replay much faster than text-model inference. TF–IDF and FLAN times were batch-amortized, whereas rules were timed per request; the figures are not equivalent online latency distributions. Task-relevant latency metrics in document streaming provide a useful direction for future interactive evaluation (H. Wang et al., 2025; H. Wang & Wang, 2026).

Table 9. Measured routing and replay time at budget four

Policy	Mean route ms	Median replay ms	95th replay ms
Fixed all	0.0000	0.0080	0.0375
Rules	0.0045	0.0245	0.0751
TF–IDF	0.0412	0.0245	0.0735
Selective TF–IDF	0.0412	0.0010	0.0261
FLAN constrained	95.3055	0.0025	0.0261
FLAN free	105.0350	0.0009	0.0011

Fixed all has no text-router inference. Replay excludes file loading, index construction, and language-model loading. Timing reflects one local run and is not a hardware-independent performance guarantee.

The benchmark has defined limits. Authored templates retain shared vocabulary; pre-resolved arguments remove extraction ambiguity. Country groups are disjoint for routing, but the predictor reuses countries across time. Soil groups lack geographic buffers, and Ethiopian groups hold out region–crop pairs rather than whole regions. Retrospective data vintages may include corrections unavailable at historical forecast time.

Retrieval also preserves source limitations: coarse grids, incomplete event catalogues, national averages, and composite indices cannot establish field truth or causality. Context bundles do not estimate flood effects. Future evaluation should test multi-turn and multilingual requests, online services, and farmer use, including whether evidence cards make uncertainty understandable (Liu & Han, 2025; Ma et al., 2025; Tu et al., 2025).

5. Conclusions

Explicit evidence contracts and verification budgets supported reliable offline execution. At four units, rules completed 100% of supported tasks, TF–IDF 95.0%, and constrained FLAN 38.6%. Removing checks enabled incorrect acceptance, while persistence outperformed learned forecasters on point error. Routing, response verification, predictive quality, and presentation therefore require separate assessment. Prospective use is needed to establish agricultural benefit.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers
Artificial Intelligence (AI) Use Disclosure: The authors declare that no artificial intelligence tools were used in the preparation of this manuscript.

References

[1] Angelopoulos, A. N., & Bates, S. (2023). Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning, 16*(4), 494–591. https://doi.org/10.1561/2200000101
[2] Chen, A., & Han, L. (2024). Small-sample tree-ensemble models for climate-smart crop selection in low-resource Ethiopian agriculture. *Journal of Advanced Computing Systems, 4*(6), 104–117. https://doi.org/10.69987/JACS.2024.40608
[3] Chen, C., Noble, I., Hellmann, J., Coffee, J., Murillo, M., & Chawla, N. (2023). *University of Notre Dame Global Adaptation Initiative: Country index technical report*. University of Notre Dame. https://gain.nd.edu/assets/522870/nd_gain_countryindextechreport_2023_01.pdf
[4] Chen, L., Zaharia, M., & Zou, J. (2024). FrugalGPT: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research*. https://openreview.net/pdf/3bc37598b9fc1a12afe49c84e43e6beb9d8e2c6a.pdf
[5] Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., … Wei, J. (2024). Scaling instruction-finetuned language models. *Journal of Machine Learning Research, 25*(70), 1–53. https://www.jmlr.org/papers/v25/23-0870.html
[6] Gao, L., Madaan, A., Zhou, S., Alon, U., Liu, P., Yang, Y., Callan, J., & Neubig, G. (2023). PAL: Program-aided language models. *Proceedings of Machine Learning Research, 202*, 10764–10799. https://proceedings.mlr.press/v202/gao23f.html
[7] Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. *Advances in Neural Information Processing Systems, 30*, 4878–4887. https://papers.nips.cc/paper_files/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html
[8] Gou, Z., Shao, Z., Gong, Y., Shen, Y., Yang, Y., Duan, N., & Chen, W. (2024). CRITIC: Large language models can self-correct with tool-interactive critiquing. In *International Conference on Learning Representations*. https://proceedings.iclr.cc/paper_files/paper/2024/hash/fef126561bbf9d4467dbb8d27334b8fe-Abstract-Conference.html
[9] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *Proceedings of Machine Learning Research, 70*, 1321–1330. https://proceedings.mlr.press/v70/guo17a.html
[10] Han, L., & Chen, A. (2024). Climate-resilience and cereal-yield joint forecasting for low- and middle-income countries with public data and grounded LLM policy explanation. *Journal of Advanced Computing Systems, 4*(1), 121–139. https://doi.org/10.69987/JACS.2024.40109
[11] Han, L., Nie, N., & Kim, P. (2026). Few-shot subnational crop portfolio ranking in Ethiopia with prior-fitted tabular models, spatial soil constraints, conformal decision sets, and evidence-locked explanations. *Smart and Intelligent Engineering Systems, 1*(1), 55–71. https://atlaspubs.com/index.php/SIES/article/view/34
[12] Han, L., & Zhou, D. (2023a). Lightweight multimodal rice leaf disease screening and Nepal-localized climate-smart advisory support: Evaluations on real rice images, Nepal SOTER soil profiles, and WorldClim climate data. *Journal of Advanced Computing Systems, 3*(4), 67–82. https://doi.org/10.69987/JACS.2023.30405
[13] Han, L., & Zhou, D. (2023b). Nepal-localized climate-smart crop advisory from crop-recommendation, cereal-yield, climate, boundary, and soil-parent data: A small-data evaluation. *Journal of Advanced Computing Systems, 3*(7), 58–73. https://doi.org/10.69987/JACS.2023.30705
[14] Han, L., & Zhou, D. (2025a). Domain-shift-aware climate-smart crop portfolio advisory for Nepal with self-supervised tabular learning and conformal shortlists. *Journal of Technology Informatics and Engineering, 4*(2), 534–558. https://doi.org/10.51903/jtie.v4i2.592
[15] Han, L., & Zhou, D. (2025b). Lightweight image–text triage for rice leaf diseases with clarify-or-abstain decisions and Nepal-context advisory support. *Journal of Technology Informatics and Engineering, 4*(1), 347–370. https://doi.org/10.51903/jtie.v4i1.593
[16] He, S., Tu, H., & Liu, I. (2023). Safe PD capacity forecasting with time-series foundation models and calibrated uncertainty for heterogeneous GPU clusters. *Journal of Advanced Computing Systems, 3*(4), 48–66. https://doi.org/10.69987/JACS.2023.30404
[17] Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns, 4*(9), Article 100804. https://doi.org/10.1016/j.patter.2023.100804
[18] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems, 33*, 9459–9474. https://papers.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[19] Li, M., Zhao, Y., Yu, B., Song, F., Li, H., Yu, H., Li, Z., Huang, F., & Li, Y. (2023). API-Bank: A comprehensive benchmark for tool-augmented LLMs. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 3102–3116). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.187

[20] Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors, 18*(8), Article 2674. https://doi.org/10.3390/s18082674

[21] Liu, E., & Han, L. (2025). From field geometry to jewelry interfaces: Multimodal AI-assisted shape translation and evidence-grounded agricultural storytelling. *International Journal of Graphic Design, 3*(1), 218–240. https://doi.org/10.51903/ijgd.v3i1.4276

[22] Ma, R., & Zhang, L. (2026). Incident-aligned probabilistic graph traffic forecasting for DOT corridor operations: Full METR-LA/PEMS-BAY benchmarking, FT-AED early event detection, and verifiable LLM-RAG briefings. *Global Sustainability & Development, 1*(2), 95–111. https://atlaspubs.com/index.php/GSD/article/view/26

[23] Ma, R., Zhang, L., & Bai, A. (2026). Cross-dataset parcel workload priors for last-mile capacity forecasting: Integrating package segmentation with delivery operation signals. *Journal of Technology Informatics and Engineering, 5*(1), 441–473. https://doi.org/10.51903/jtie.v5i1.564

[24] Ma, R., Zhang, L., & Song, T. (2025). Computer-vision-informed visual explanation cards for autonomous-driving traffic-sign alerts: Localization, classification, and retrieved evidence on GTSDB. *International Journal of Graphic Design, 3*(2), 455–473. https://doi.org/10.51903/ijgd.v3i2.3991

[25] Ong, I., Almahairi, A., Wu, V., Chiang, W.-L., Wu, T., Gonzalez, J. E., Kadous, M. W., & Stoica, I. (2024). *RouteLLM: Learning to route LLMs with preference data* (Version 1) [Preprint]. arXiv. https://arxiv.org/abs/2406.18665v1

[26] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*(85), 2825–2830. https://www.jmlr.org/papers/v12/pedregosa11a.html

[27] Poggio, L., de Sousa, L. M., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Ribeiro, E., & Rossiter, D. (2021). SoilGrids 2.0: Producing soil information for the globe with quantified spatial uncertainty. *SOIL, 7*(1), 217–240. https://doi.org/10.5194/soil-7-217-2021

[28] Qin, Y., Liang, S., Ye, Y., Zhu, K., Yan, L., Lu, Y., Lin, Y., Cong, X., Tang, X., Qian, B., Zhao, S., Hong, L., Tian, R., Xie, R., Zhou, J., Gerstein, M., Li, D., Liu, Z., & Sun, M. (2024). ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In *International Conference on Learning Representations*. https://proceedings.iclr.cc/paper_files/paper/2024/hash/28e50ee5b72e90b50e7196fde8ea260e-Abstract-Conference.html

[29] Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography, 40*(8), 913–929. https://doi.org/10.1111/ecog.02881

[30] Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2023). Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems, 36*, 68539–68551. https://doi.org/10.52202/075280-2997

[31] Tellman, B., Sullivan, J. A., Kuhn, C., Kettner, A. J., Doyle, C. S., Brakenridge, G. R., Erickson, T. A., & Slayback, D. A. (2021). Satellite imaging reveals increased proportion of population exposed to floods. *Nature, 596*(7870), 80–86. https://doi.org/10.1038/s41586-021-03695-w

[32] Tu, H., Zhao, S., & He, S. (2024). LLM-augmented salable GPU supply forecasting for disaggregated recommendation serving: Predicting instance readiness, scheduling delay, and capacity risk. *Journal of Advanced Computing Systems, 4*(9), 85–102. https://doi.org/10.69987/JACS.2024.40908

[33] Tu, H., Zhao, S., & Zhou, A. (2025). Visual brief cards for advertising design: A structured UI/UX framework for turning creative intentions into graphic design decisions. *International Journal of Graphic Design, 3*(1), 210–226. https://doi.org/10.51903/ijgd.v3i1.3714

[34] Wang, E., & Wang, H. (2023). QoE-driven reinforcement learning for joint bitrate, rebuffering, and TTFF optimization in HLS/DASH. *Journal of Advanced Computing Systems, 3*(2), 50–59. https://doi.org/10.69987/JACS.2023.30204

[35] Wang, H., Li, S., Li, C., Wang, Z., & Peng, H. (2020). Low complexity hybrid beamforming methods for single-user millimeter wave MIMO systems. *The Frontiers of Society, Science and Technology, 2*(12), 16–29. https://francis-press.com/papers/2548

[36] Wang, H., Ren, Y., & Chang, X. (2025). Layout-aware progressive PDF rendering: AI prioritization of PDF slices to reduce time-to-functional-first-frame on FUNSD. *Journal of Technology Informatics and Engineering, 4*(2), 425–446. https://doi.org/10.51903/jtie.v4i2.523

[37] Wang, H., & Wang, S. (2026). Query-conditioned document-VLM streaming for time-to-first-correct-answer: Reinforcement learning over progressive PDF tiles. *Global Sustainability & Development, 1*(2), 65–81. https://atlaspubs.com/index.php/GSD/article/view/23

[38] Wang, H., & Wu, P. (2024). Content-adaptive codec selection between AV1 and H.264: A controlled case study of bandwidth savings versus latency at matched quality. *Journal of Advanced Computing Systems, 4*(7), 83–92. https://doi.org/10.69987/JACS.2024.40707

[39] Wen, Z. (2025). Research on a deep learning-based dynamic credit risk assessment model for enterprises. In *Proceedings of the 2024 5th International Conference on Big Data Economy and Information Management* (pp. 389–393). Association for Computing Machinery. https://doi.org/10.1145/3724154.3724218

[40] Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*. https://arxiv.org/abs/2210.03629

[41] Zhang, L., & Ma, R. (2026). Calibrated spatiotemporal GraphRAG for urban road safety: Network-constrained severe-crash hotspot forecasting, evidence-grounded countermeasure ranking, and risk-controlled abstention. *Smart and Intelligent Engineering Systems, 1*(1), 36–54. https://atlaspubs.com/index.php/SIES/article/view/30

[42] Zhang, L., Ma, R., & Greg, P. (2025). Digital-twin dispatching for urban mobility via spatio-temporal transformers and offline reinforcement learning. *Journal of Technology Informatics and Engineering, 4*(2), 337–363. https://doi.org/10.51903/jtie.v4i2.501

[43] Zhang, L., & Zhang, E. (2024). LLM-explained graph traffic forecasting for DOT corridor operations: Full empirical evaluation on METR-LA and PEMS-BAY with crash evidence. *Journal of Advanced Computing Systems, 4*(5), 135–145. https://doi.org/10.69987/JACS.2024.40510

[44] Zhong, Z. S., Zhang, L., & Ma, D. (2023). Spatial RAG for urban crash hotspot discovery and safety countermeasure recommendation. *Journal of Advanced Computing Systems, 3*(12), 45–61. https://doi.org/10.69987/JACS.2023.31206

[45] Zhou, B., Wang, H., & Chang, X. (2025). Distilling VMAF into an edge-deployable quality predictor: A pilot shot-level proxy with LLM-ready quality tokens. *Journal of Technology Informatics and Engineering, 4*(2), 447–463. https://doi.org/10.51903/jtie.v4i2.522

[46] Zhou, H., & Han, L. (2026). Flood-shock-aware climate-resilience and cereal-yield forecasting for low- and middle-income countries with cross-series pretrained time-series models, conformal downside risk, and evidence-locked LLM-compatible policy explanations. *Global Sustainability & Development, 1*(2), 124–141. https://atlaspubs.com/index.php/GSD/article/view/32