Stout in Agriculture and Animal Science (SAAS)
ISSN: 2979-2835
DOI: https://doi.org/10.61424/g6rbs789
Journal Homepage: https://stoutjournals.org/index.php/SAAS

Capability-Profiled Agricultural Language-Model Systems: Calibration and Cost-Aware Routing across Memorization, Understanding, Reasoning, and Generation

Sylvia He
Computer Science, University of Pittsburgh, Pittsburgh, PA, USA
Corresponding Author: Sylvia He, **E-mail**: s.he_pitt@outlook.com

ARTICLE INFORMATION

Submitted: March 10, 2026
Accepted: June 19, 2026
Published: July 31, 2026
Volume: 2
Issue: 1

KEYWORDS

AgriEval; agricultural question answering; capability profiling; calibration; cost-aware routing; retrieval; self-verification; Qwen2.5

ABSTRACT

Agricultural language systems must span factual recall, diagnosis, numerical reasoning, and management advice, yet a single score conceals these differences. This study profiles the AgriEval 2025 snapshot using 14,697 multiple-choice and 2,167 open-answer questions across six domains, 29 subdomains, 15 tasks, and four capability families. Duplicate-normalized question groups were assigned to five stratified folds before training, calibration, and held-out testing. Experiments compared Qwen2.5-0.5B-Instruct in direct and self-verification conditions with position, character TF-IDF, case-retrieval, fusion, and answer-copy baselines. On 2,940 multiple-choice test questions, CharPair achieved the highest exact accuracy, 0.3833 (95% confidence interval [0.3653, 0.3997]); Qwen direct reached 0.3286 [0.3119, 0.3459]. Isotonic regression reduced Qwen's 10-bin expected calibration error from 0.2646 to 0.0147 without changing answers. Self-verification reduced accuracy to 0.2922: 261 errors were corrected, but 368 correct direct answers were corrupted. A calibration-trained CharPair-to-Qwen router sent 9.01% of questions to Qwen, attained 0.3810 accuracy, and required 60.455 ms per question, compared with 0.3833 and 0.549 ms for CharPair alone. On 433 open-answer items, Qwen obtained character ROUGE-L of 0.1555 [0.1477, 0.1636], while VerifiedRAG reached 0.1641. The experiments show that calibration can improve probability reliability even when retrieval, verification, and routing do not improve answer quality. Capability, subgroup, uncertainty, and measured-compute profiles therefore provide a more informative assessment than one leaderboard score.

1. Introduction

Agricultural decision support spans disease recall, concept discrimination, symptom-based diagnosis, input-rate calculation, and management advice. These tasks differ in knowledge demand, error severity, and justified computation. One aggregate score obscures weak subfields, unreliable confidence, and unnecessary compute.

Massive Multitask Language Understanding established broad knowledge testing (Hendrycks et al., 2021), while C-Eval and CMMLU adapted it to Chinese settings (Huang et al., 2023; H. Li et al., 2024). AgriEval organizes Chinese agricultural questions by domain, format, cognition, and task (Yan et al., 2026), reflecting wider interest in systems that connect technical knowledge with farm advice (H. Li et al., 2025; Tzachor et al., 2023).

Three gaps remain: dominant domains can drive aggregate performance; accuracy and confidence are distinct; and routing is an economic as well as predictive decision. Miscalibration matters when confidence controls review or abstention (Guo et al., 2017). The study evaluates a 0.49-billion-parameter instruction model alongside interpretable reference systems. Character n-gram classifiers measure local lexical structure, nearest neighbors measure case similarity, calibration maps scores to exact-correctness probabilities, and routing relates quality to measured compute. Open-answer copying is reported separately from generated responses.

The contributions are a group-aware audit and split of 16,864 questions; held-out evaluation of Qwen direct inference, self-verification, six multiple-choice baselines, four answer-copy baselines, and generation; and reliability, risk–coverage, uncertainty, and cost analyses.

The research questions are:

1. How does a small instruction model compare with heuristic, character-based, and case-retrieval baselines on AgriEval?
2. How strongly do performance differences depend on cognitive family, question format, agricultural domain, and subdomain?
3. To what extent can post-hoc calibration and confidence ranking support reliable selective use?
4. Do self-verification and calibration-trained routing improve quality enough to justify their measured compute?

These questions define reference expectations for accuracy, subgroup stability, calibration, and compute efficiency.

2. Literature Review

2.1 Agricultural language-model evaluation

Agricultural language systems connect technical literature, field observations, extension knowledge, and farmer questions. Reviews describe rapid growth but uneven evaluation depth (H. Li et al., 2025; Shaikh et al., 2025). Their extension-service potential comes with reliability and access risks (De Clercq et al., 2024; Tzachor et al., 2023).

AgriEval links question formats with cognitive labels and agricultural taxonomy (Yan et al., 2026), enabling separate analysis of memorization, understanding, reasoning, and generation. The design follows the principle that correct responses can reflect different cognitive processes (Bloom et al., 1956).

Direct agricultural evidence remains the center of this study. Cross-domain work is used below only when it instantiates an evaluation mechanism also examined here: capability decomposition, retrieval and provenance, probability calibration, selective prediction, self-verification, routing, or human-facing uncertainty communication. Results from medicine, finance, security, education, and cloud operations are not treated as evidence of agricultural accuracy.

2.2 Capability profiling and cross-domain transfer

Capability profiles are useful when they separate task families, transfer conditions, uncertainty, and resource burden rather than reducing them to one score. Parameter-efficient medical pretraining has shown task-dependent rather than uniform transfer (Mi et al., 2025). Subject-independent brain-computer-interface evaluation and cross-dataset wearable activity transfer similarly distinguish within-domain, leave-one-subject-out, and adaptation conditions (Wu et al., 2025; J. Zhang, 2025). The approximately shared-features framework provides a formal account of when transfer may remain possible despite domain mismatch (Zhong, Pan, et al., 2025). These studies motivate reporting AgriEval performance by cognition and agricultural subdomain instead of assuming that pooled success transfers uniformly.

Compact-system studies also show why deployability belongs beside predictive quality. Retrieval-summary integration has been evaluated for lightweight code intelligence, VMAF has been distilled into an edge quality proxy, and self-supervised transformer representations have been tested for log anomalies (Y. Li, 2024; Xin, 2026i; B. Zhou, Wang, et al., 2025). Rubric-grounded essay scoring, classroom early-warning analytics, and teacher-facing course explanations further illustrate that a benchmark score becomes operationally meaningful only when its decision target and review interface are explicit (Xin, 2026a, 2026b; J. Zhang, 2026). Finally, rank-aggregation theory and uncertainty analysis of spectral estimators caution against treating a single observed ordering as exact (Zhong & Ling, 2024a, 2024b).

2.3 Retrieval, provenance, and evidence interfaces

Retrieval-augmented language systems condition a model on retrieved material (Guu et al., 2020; Lewis et al., 2020), and dense passage retrieval learns representations for open-domain search (Karpukhin et al., 2020). Here, "CaseRAG" and "HybridRAG" denote non-LLM case retrieval or score fusion, without external evidence or response synthesis.

TF-IDF supports cosine comparison between sparse vectors (Salton & Buckley, 1988), and nearest neighbors transfer labels from similar cases (Cover & Hart, 1967). Character n-grams avoid fixed Chinese segmentation but can miss negation, causality, and contextual mismatch.

Applied RAG research increasingly separates retrieval quality, answer quality, and evidence traceability. Private-document DevOps question answering, bilingual incident triage, and word-level evidence-chain analysis all emphasize grounding and source attribution (C. Li et al., 2024; Liu et al., 2024; B. Zhang et al., 2024). Related work has examined code search, accounting-aware due diligence, and budgeted multi-hop retrieval (Y. Chen et al., 2025; Y. Li, 2024; Su et al., 2025). Unanswerable-question abstention and deterministic provenance explanation further separate missing evidence from unsupported generation (Jin, 2025a; Zhong, Chen, et al., 2025). These distinctions are important here because training-case similarity is not equivalent to retrieving an authoritative agronomic source.

The same decomposition appears in scientific, financial, climate, multilingual, and legal settings. Evidence ranking has been studied for ClimateCheck claims, hybrid retrieval and listwise reranking for FiQA, multilingual RAG for migration information, and regulation-aware retrieval for product counsel (Y. Chen & Xu, 2026; J. Li & Zhou, 2026; Meng et al., 2026; Su, Chen, et al., 2026). Scientific-search and accounting-disclosure cards add visible evidence hierarchy, while counseling recommendation cards connect retrieved support to a human professional rather than treating the generated response as an autonomous decision (Jin, 2025b; K. Zhang et al., 2025; Y. Zhang & Zhang, 2025b).

Operational studies further demonstrate why provenance and command safety should be evaluated separately from fluent text. Retrieved normal prototypes have been used to explain OpenStack anomalies; HDFS failure narratives, cloud troubleshooting runbooks, and routed AIOps summaries have been constrained by retrieved logs or documents (C. Li et al., 2026; Sun et al., 2026; Xin, 2025a; B. Zhang et al., 2026). Incident-visualization cards and accounting-aware agents extend the same principle to human review of distributed logs and tokenized-asset operations (Nie et al., 2026; S. Zhou et al., 2026). Predicting retrieval quality itself is a further safeguard when the retriever may fail before generation begins (R. Zhang et al., 2025).

2.4 Calibration, selective prediction, and uncertainty

Calibration asks whether cases assigned confidence (p) are correct at approximately rate (p). Platt scaling fits a sigmoid (Platt, 2000), whereas isotonic regression learns a monotonic nonparametric mapping (Zadrozny & Elkan, 2002). Neural systems can be substantially miscalibrated before such adjustment (Guo et al., 2017).

ECE summarizes bin-level confidence–accuracy gaps, Brier score is squared probability error, and log loss penalizes confident mistakes (Pedregosa et al., 2011). Calibration here targets binary exact correctness.

Selective prediction answers only cases above a threshold and defers the rest; risk–coverage analysis measures how error changes with coverage (Geifman & El-Yaniv, 2017). Calibration does not guarantee strong ranking, so lower-coverage accuracy and residual risk must both be measured.

Cross-domain empirical work reinforces the difference between discrimination and probability quality. Credit-default tiering and model-risk workflows combine calibration with rejection or distribution-free uncertainty; fraud and bankruptcy studies add extreme-imbalance analysis where rare errors dominate practical value (Y. Chen et al., 2023; Y. Chen et al., 2024; Jin et al., 2024a, 2024b). Noisy-neighbor VM degradation modeling provides a related example in which residual risk, rather than mean fit alone, drives the operational question (Nie & Zheng, 2024).

Calibration has also been paired with learned fusion, medical explanation, and high-stakes review. Uncertainty-aware LiDAR-camera fusion reports probability reliability separately from detection, while object-level tracking combines learned perception with explicit association and filtering (Xin, 2025b, 2026e). Medical image explanation cards and patient-facing safety-card frameworks tie confidence to constrained communication (C. Li et al., 2025; Zhong, Wu, et al., 2025; B. Zhou, Li, et al., 2025). Financial dashboards and AI-infrastructure brief cards similarly translate risk estimates into reviewable visual hierarchy rather than presenting a scalar without context (Z. Li et al., 2025; Liu et al., 2025).

Forecasting and resource studies broaden the uncertainty question from class probabilities to future demand and operational envelopes. Calibrated capacity forecasting, cross-cloud transfer, and few-shot tenant forecasting assess uncertainty under workload or topology shift (He et al., 2023, 2024, 2025). Conformal oversubscription, multi-horizon GPU forecasts, and power-aware inventory planning connect those forecasts to resource decisions (S. Chen et al., 2024; He et al., 2026; Zhao et al., 2025). News-macro fusion and probabilistic bike-demand forecasting make the same distinction for economic volatility and mobility demand (Xin, 2026g; H. Zhou & Zhang, 2025). These applications support the present choice to report ECE, Brier score, risk-coverage, and subgroup stability rather than interpreting accuracy as complete reliability.

Selective action requires a valid downstream policy. Profit-aware GPU admission, conformal alert control, counterfactual credit explanations, and agent-trajectory reliability prediction all connect uncertainty to escalation, rejection, or review (Xin, 2026c,

2026f; Zhao et al., 2026; Zhong et al., 2026). The present router is therefore evaluated as a held-out policy with measured cost, not inferred to be useful merely because its component confidences were calibrated.

2.5 Safety, verification, and attack-aware evaluation

Self-verification can help only when it corrects more errors than it introduces. Evidence-based toxicity checks and staged refusal systems explicitly measure both attack resistance and utility preservation, while continual red-teaming treats guardrails as components that must adapt to new jailbreak behavior (D. Zheng & Li, 2024; D. Zheng et al., 2023; D. Zheng et al., 2024). These designs motivate the paired correct-to-wrong and wrong-to-correct analysis used for the Qwen critique pass.

Security studies also show that a concise explanation does not substitute for attack detection. Multi-source root-cause attribution, TinyLLM-assisted intrusion detection, and language-guided DDoS feature selection each retain an explicit detector before narrative generation (Y. Li & Lu, 2025; Liu et al., 2023; Lu & Zhou, 2024). Sequence-level CloudTrail attack-chain models and host-system-call localization preserve the same separation (Bai, Chen, et al., 2026; Xin, 2026d). Numerical-reasoning guardrails and privacy/data-integrity risk cards likewise place deterministic checks and human oversight around LLM outputs (Z. Li et al., 2026; Su, Rao, et al., 2026). Agricultural self-verification should therefore be judged by changed outcomes and safety-relevant failure patterns, not by the apparent plausibility of the second response.

2.6 Small models, routing, and measured cost

Qwen2.5 includes a 0.49-billion-parameter instruction model with Chinese generation (Qwen Team, 2025). A critique pass helps only when corrections exceed introduced errors; routing helps only when quality gains offset escalation cost. FrugalGPT and RouteLLM formalize this cascade tradeoff (L. Chen et al., 2024; Ong et al., 2025). Decoding, calibration, and routing decisions were therefore selected on validation data.

Resource-aware work offers several analogues for this evaluation. GPU demand and inventory studies route or admit workloads under forecast risk, while cost-aware AIOps routing decides when more expensive analysis is justified (He et al., 2026; C. Li et al., 2026; Zhao et al., 2025, 2026). Capacity dashboards make the predicted risk and operational consequence visible to a human reviewer (Liu et al., 2025). Offline counterfactual evaluation provides a complementary warning: a policy that looks attractive under weak support can be misranked unless overlap and uncertainty are checked (Mu et al., 2025).

The same cost-quality tension appears in foundation-model forecasting, cross-cloud transfer, and compact edge predictors (He et al., 2023, 2024, 2025; B. Zhou, Wang, et al., 2025). In the present study, measured CPU time and milliseconds per correct answer are therefore treated as outcomes. A routed agricultural system is useful only if its escalated model adds held-out quality commensurate with that cost.

2.7 Personalization and human-facing support

Agricultural advice is often user- and context-dependent, but personalization can preserve stale or unsupported information. Confidence-gated memory writing and time-aware retrieval address update and forgetting across dialogue sessions, while federated preference learning addresses privacy in knowledge-grounded chat (Kuo et al., 2025; Sun et al., 2023). Customer representation learning and review-grounded recommendation further show how behavioral or textual evidence can support segmentation and explanation (Chang et al., 2026; Xin, 2026h).

Human-facing cards are a recurring interface pattern rather than a claim that presentation fixes model error. E-commerce recommendation cards expose evidence and confidence; therapist-facing retrieval and strategy-aware response control keep a professional or explicit response policy in the loop (B. Zhang et al., 2025; Y. Zhang & Zhang, 2025a; Y. Zhang & Zhou, 2026). Uplift modeling for email advertising separates predicted response from the policy decision of whom to target and with which creative, while privacy-robust incrementality work tests how that policy problem changes when user identifiers are unavailable (Bai, Wang, et al., 2026; Mu et al., 2023). For agricultural generation, these precedents support future evidence-linked expert review, but the present experiments stop at measured retrieval and generation performance.

2.8 Evaluation metrics and statistical inference

Multiple-choice evaluation permits exact set matching across true/false, single-choice, and multiple-selection formats. Option micro-F1 gives partial credit, while macro-F1 over answer patterns reveals frequency effects.

Open-answer evaluation uses ROUGE-L longest-common-subsequence agreement (Lin, 2004), character n-gram chrF (Popović, 2015), and TF-IDF cosine. These quantify reference overlap, not agronomic truth.

Uncertainty around test estimates was computed by resampling test items. Bootstrap intervals are useful when an analytic sampling distribution is inconvenient (Efron, 1979). McNemar's test addresses paired binary outcomes on the same cases (McNemar, 1947).

3. Methodology

3.1 Dataset and integrity audit

AgriEval contained 14,697 multiple-choice and 2,167 open-answer records across six domains, 29 subdomains, 15 tasks, and four capabilities. Normalized stems retained case, whitespace, and punctuation variants within one split group. Current fields governed conflicts with deprecated fields; all answer labels were valid.

Table 1. Dataset integrity audit.

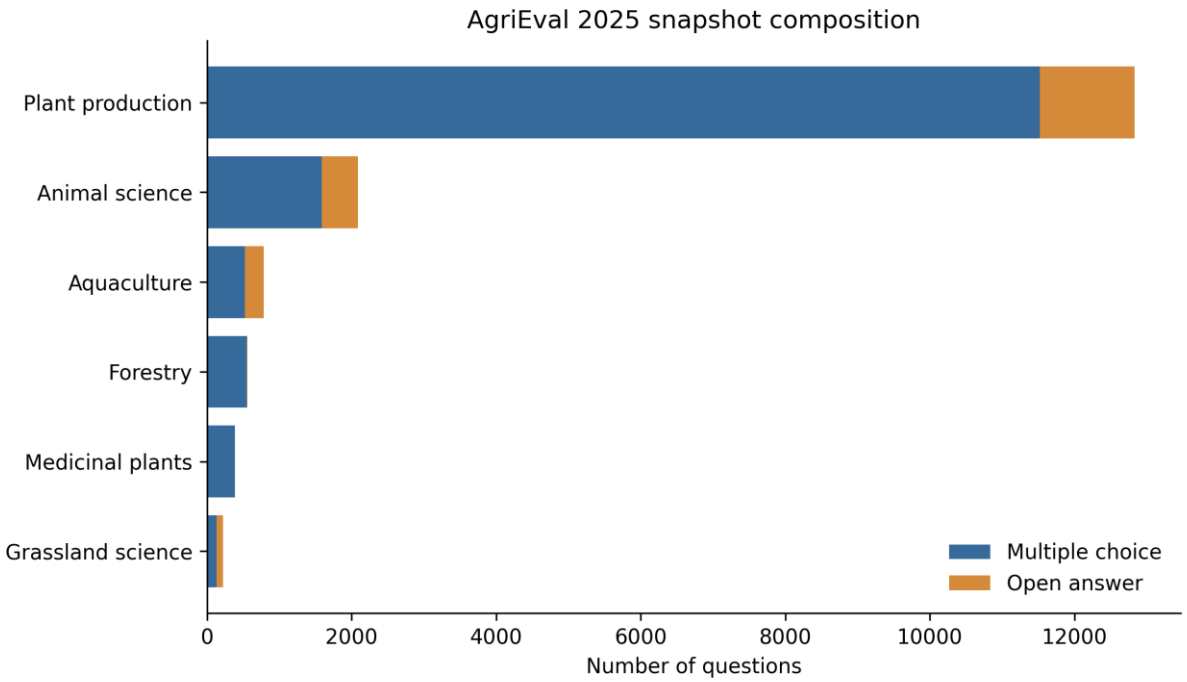
Check	Multiple choice	Open answer	Overall
Parsed rows	14,697	2,167	16,864
Rows in normalized duplicate-stem groups	791	50	841
Normalized duplicate groups, multiple choice	388	—	388
Cross-file normalized stems	—	—	9
Missing question-format labels	0	63	63
Deprecated/current field conflicts	2,774	616	3,390
Invalid answer-option sets	0	—	0

Plant production contributed 76.08% of records, making domain-specific reporting necessary.

Table 2. Domain composition of the complete corpus.

Domain	Multiple choice	Open answer	Total	Share (%)
Medicinal plants	387	1	388	2.30
Animal science	1,586	500	2,086	12.37
Forestry	548	8	556	3.30
Plant production	11,520	1,310	12,830	76.08
Aquaculture	525	261	786	4.66
Grassland	131	87	218	1.29

Figure 1. Domain composition by question format. Counts are corpus records, not reweighted estimates.



Memorization contained 7,198 records, understanding 5,538, reasoning 1,961, and generation 2,167; generation corresponded to open answers.

Table 3. Cognitive-task composition.

Capability	Task	Count
Memorization	Operation rules	116
Memorization	Terminology	125
Memorization	Fundamental concepts	6,077
Memorization	Production-management essentials	880
Understanding	Knowledge verification	1,961
Understanding	Type identification	2,253
Understanding	Key-point summary	1,324
Reasoning	Numerical reasoning	707
Reasoning	Production-management reasoning	471
Reasoning	Growth-status reasoning	273
Reasoning	Disease diagnosis	403
Reasoning	Genetic reasoning	107
Generation	Causal question answering	142
Generation	Knowledge question answering	1,700
Generation	Strategy question answering	325

3.2 Grouped data splitting

Five grouped stratified folds used seed 20250729. Multiple-choice strata combined format, cognition, and domain; open strata used task. Folds 2–4 were training, fold 1 validation, and fold 0 test. No normalized stem crossed folds.

Table 4. Group-aware split statistics.

Format	Split	Records	Unique normalized stems	Domains	Subdomains
Multiple choice	Train	8,817	8,576	6	29
Multiple choice	Validation	2,940	2,859	6	29
Multiple choice	Test	2,940	2,859	6	29
Open answer	Train	1,300	1,285	6	27
Open answer	Validation	434	429	5	23
Open answer	Test	433	428	5	23

3.3 Multiple-choice systems

PositionPrior was an add-one-smoothed answer-position heuristic conditioned on format and option count; LongestOption selected the longest normalized option.

CharPair encoded question–option pairs with character TF-IDF: 2–5-grams, document frequency 2–0.998, sublinear term frequency, and 70,000 features. A class-balanced averaged stochastic-gradient logistic classifier used log loss, L2 regularization, (^{-5}), and at most 120 iterations (Bottou, 2010). Its short view used 1–2-grams, 45,000 features, and (^{-5}).

CaseRAG retrieved seven training questions under 2–5-character TF-IDF cosine and scored options by 1–4-character similarity to their correct-answer text. It did not retrieve external evidence.

HybridRAG fused CharPair/CaseRAG scores at 0.70/0.30; VerifiedHybrid fused that result with the short view at 0.80/0.20. “Verified” denotes dual-view scoring. Validation selected multiple-answer thresholds from 0.20–0.80 with at least two options; other formats used argmax.

3.4 Open-answer systems

Four open-answer baselines copied one training answer. CharRAG used 2–5-character retrieval; DomainRAG restricted candidates by domain and blended long/short similarity at 0.70/0.30; HybridRAG removed the restriction; VerifiedRAG used the long-view winner on agreement and the domain-hybrid winner otherwise.

3.5 Local language-model inference and self-verification

Qwen2.5-0.5B-Instruct ran locally as a Q4_K_M GGUF checkpoint (SHA-256 74a4da8c9fdbcd15bd1f6d01d621410d31c6fc00986f5eb687824e7b93d7a9db) through node-llama-cpp 3.19.1. Prompts presented each Chinese question and labeled options. Next-token probabilities were restricted to valid labels A–G. A stratified 512-item validation sample selected absolute probability 0.02 for multiple answers; other formats used argmax.

Self-verification added Qwen’s direct answer and an instruction to recheck it. Its decoder and calibration were fitted independently on the same validation sample. Open generation used a 768-token context, greedy decoding, and at most 64 tokens; the Chinese prompt requested a concise answer without restating the question. Reference answers never entered prompts.

3.6 Calibration and decision analyses

Baseline confidence combined top probability and top-two margin for single-choice and true/false items and averaged option certainty for multiple selection. Raw, Platt, and isotonic mappings were fitted to validation exact correctness and selected by ECE, then Brier score.

For Qwen, raw confidence was the maximum valid-label probability or the selected-label mass for multiple answers. Isotonic regression mapped it to exact-correctness probability on the 512 validation items; risk–coverage ranked test items by that probability.

The router ran CharPair first and sent confidence below a threshold to Qwen. On the 512 validation items, 0.1951 was the least costly point within 0.005 accuracy of the best route. Test latency included CharPair throughout and Qwen when escalated.

3.7 Metrics, latency, and statistical analysis

Multiple-choice metrics were exact accuracy, option micro-F1, answer-pattern macro-F1, Brier, 10-bin ECE, NLL, AURC, latency, and milliseconds per correct answer; multiple-answer results added Jaccard. Open metrics were character ROUGE-L, chrF, TF-IDF cosine, and the ROUGE-L ≥ 0.25 rate.

After warm-up, CharPair used two timing repetitions and CaseRAG one end-to-end run. Open retrieval timing excluded index preparation. Qwen effective latency divided summed chunk wall time by evaluated items; self-verification included both passes.

Baseline intervals used 1,000 item bootstraps and Qwen intervals 5,000; paired differences resampled common items. Exact McNemar tests compared discrete outcomes. The Linux host exposed nine AMD EPYC 9V74 vCPUs and 21 GiB memory; Qwen used two threads with 16 MCQ or eight open sequences. Python 3.12.13, NumPy 2.3.5, and pandas 2.2.3 were used.

Table 5. Core experimental configuration.

Component	Fixed configuration
Split	Five grouped stratified folds; train = folds 2–4, validation = 1, test = 0
CharPair	Character TF-IDF 2–5 grams; 70,000 features; balanced averaged SGD logistic model
CaseRAG	Training-only seven-nearest-question index; option-to-retrieved-answer character similarity
VerifiedHybrid	CharPair/CaseRAG validation blend 0.70/0.30; final hybrid/short-view blend 0.80/0.20
Calibration	Raw, Platt, and isotonic mappings fitted on validation correctness; selected by ECE then Brier
Qwen MCQ	Qwen2.5-0.5B-Instruct Q4_K_M; forced valid-label scoring; 512-token context; 16 sequences; 2 CPU threads
Self-verification	Direct decoded answer added to a second checking prompt; total cost includes both passes
Cost router	CharPair first; escalate below validation-selected calibrated-confidence threshold 0.1951
Open answer	Four training-case retrieval systems plus Qwen greedy generation; 64-token output limit
Random seed	20250729

4. Results and Discussion

4.1 Overall multiple-choice performance

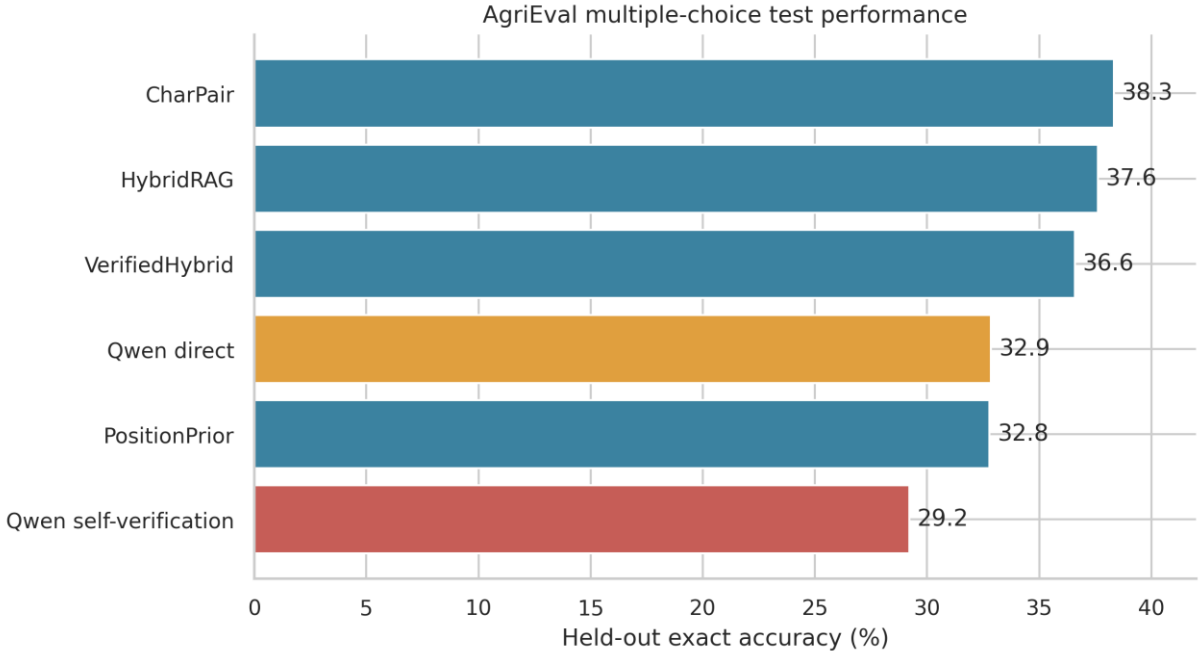
CharPair led the 2,940-item test set at 0.3833 exact accuracy; HybridRAG followed at 0.3762. Qwen direct reached 0.3286 and self-verification 0.2922. Qwen trailed CharPair by 5.48 points (paired 95% interval [−0.0769, −0.0316]; McNemar ($p < .001$)).

Table 6. Multiple-choice performance, calibration, and scoring cost on the test set.

System	Exact accuracy	Option micro-F1	Answer macro-F1	Brier	ECE-10	NLL	Latency (ms/q)	ms/correct
Position Prior	0.3279	0.4895	0.0478	0.2189	0.0105	0.6293	0.0030	0.0091
Longest Option	0.2714	0.4793	0.0808	0.1982	0.0175	0.5859	0.0050	0.0184
CharPair	0.3833	0.5580	0.0724	0.2296	0.0116	0.6528	0.5491	1.4324
CaseRAG	0.2741	0.3937	0.0679	0.1991	0.0345	0.5882	2.8864	10.5287
HybridRAG	0.3762	0.5387	0.0831	0.2319	0.0201	0.6608	3.4355	9.1324
Verified Hybrid	0.3660	0.5240	0.0669	0.2288	0.0303	0.6524	3.6251	9.9050
Qwen direct	0.3286	0.5099	—	0.2190	0.0147	0.6447	664.6151	2,022.7415
Qwen self-verification	0.2922	0.4561	—	0.2062	0.0340	0.6099	1,198.7372	4,102.7793

Note. The first six systems are non-LLM baselines. Qwen self-verification latency includes its direct and second passes. Calibration metrics use validation-fitted mappings.

Figure 2. Held-out exact accuracy for the principal multiple-choice conditions.



PositionPrior's 0.3279 reflects format structure rather than agricultural competence; option-permutation checks remain important for language models (Pezeshkpour & Hruschka, 2024; C. Zheng et al., 2024). CharPair used 0.549 ms per question versus Qwen's 664.615 ms, yet was more accurate.

4.2 Cognitive, format, domain, and subdomain profiles

CharPair led understanding (0.3917) and memorization (0.3864); HybridRAG led reasoning (0.3664), 1.78 points above CharPair.

Table 7. Exact multiple-choice accuracy by cognitive family.

System	Reasoning (n = 393)	Understanding (n = 1,108)	Memorization (n = 1,439)
PositionPrior	0.2774	0.3592	0.3176
LongestOption	0.2341	0.3069	0.2543
CharPair	0.3486	0.3917	0.3864
CaseRAG	0.2723	0.3321	0.2300
HybridRAG	0.3664	0.3845	0.3725
VerifiedHybrid	0.3461	0.3764	0.3634
Qwen direct	0.2392	0.3745	0.3176
Qwen self-verification	0.2265	0.3439	0.2703

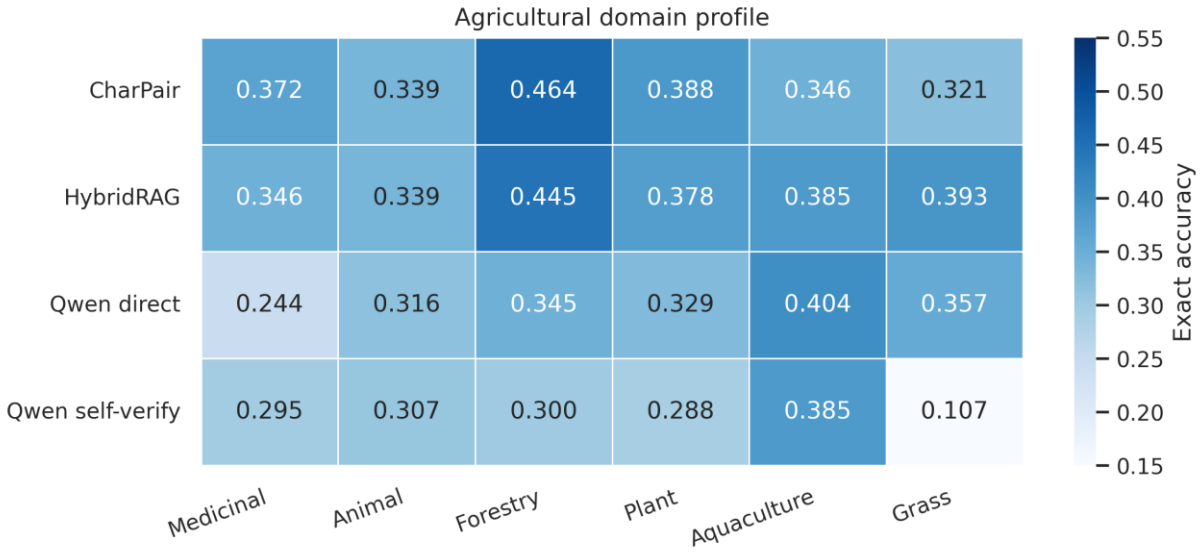
Qwen direct was strongest on understanding (0.3745) and weakest on reasoning (0.2392); verification reduced all three scores. CaseRAG led true/false, HybridRAG single choice, and CharPair multiple selection. Qwen's multiple-selection option micro-F1 of 0.8362 far exceeded 0.2865 exact-set accuracy, indicating extra or missing labels.

Table 8. Selected format and domain results.

Group	n	PositionPrior	LongestOption	CharPair	CaseRAG	HybridRAG	VerifiedHybrid
True/false	397	0.5013	0.4987	0.5038	0.5516	0.5038	0.5038
Single choice	2,159	0.3043	0.2177	0.3678	0.2640	0.3840	0.3807
Multiple selection	384	0.2813	0.3385	0.3464	0.0443	0.2005	0.1406
Medicinal plants	78	0.2949	0.2564	0.3718	0.2051	0.3462	0.3462
Animal science	316	0.3418	0.2468	0.3386	0.2943	0.3386	0.3513
Forestry	110	0.3818	0.2727	0.4636	0.3182	0.4455	0.4455
Plant production	2,304	0.3242	0.2782	0.3885	0.2708	0.3785	0.3633
Aquaculture	104	0.3654	0.2596	0.3462	0.2692	0.3846	0.4038
Grassland	28	0.2143	0.0714	0.3214	0.3571	0.3929	0.3571

Leaders varied by domain. Plant production had 2,304 test cases but grassland only 28, where one answer changes accuracy by 3.57 points.

Figure 3. Exact-accuracy profile across the six major agricultural domains for the principal sparse, retrieval, direct, and self-verification conditions.



Subdomain results varied further: forest protection favored CharPair, while weed science and aquaculture favored fusion. Very small strata support description, not stable ranking.

Table 9. Selected supported-subdomain exact accuracy.

Subdomain	n	CharPair	CaseRAG	HybridRAG	VerifiedHybrid
Forest protection	56	0.6071	0.3750	0.5536	0.5714
Vegetable science	46	0.5217	0.2826	0.5000	0.4348
Weed science	68	0.4118	0.2794	0.5294	0.5294
Aquaculture	49	0.3673	0.2653	0.4286	0.4490
Pesticides	48	0.2500	0.1458	0.2917	0.2292

HybridRAG is the reasoning reference, CharPair the pooled reference, and no system is satisfactory for multiple selection.

4.3 Calibration and reliability

Calibration improved every system. Isotonic regression reduced CharPair ECE from 0.2017 to 0.0116, Brier from 0.2801 to 0.2296, and NLL from 0.7885 to 0.6528. Selected ECE ranged from 0.0105 to 0.0345.

Table 10. Raw and selected multiple-choice calibration.

System	Raw Brier	Raw ECE-10	Raw NLL	Selected mapping	Selected Brier	Selected ECE-10	Selected NLL
PositionPrior	0.2666	0.1706	0.7448	Platt	0.2189	0.0105	0.6293
LongestOption	0.3064	0.2808	1.0327	Platt	0.1982	0.0175	0.5859
CharPair	0.2801	0.2017	0.7885	Isotonic	0.2296	0.0116	0.6528
CaseRAG	0.3472	0.3127	1.1635	Isotonic	0.1991	0.0345	0.5882
HybridRAG	0.2833	0.1975	0.7843	Isotonic	0.2319	0.0201	0.6608
VerifiedHybrid	0.3031	0.2320	0.8483	Isotonic	0.2288	0.0303	0.6524

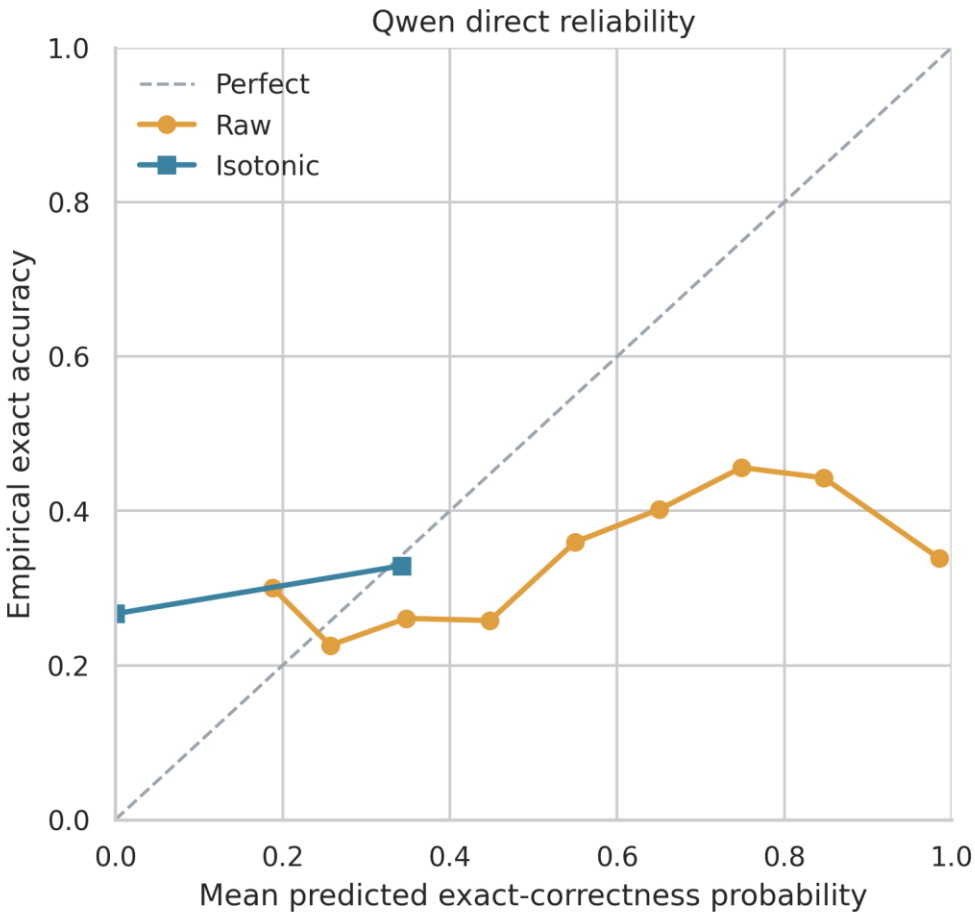
Calibration left predictions unchanged. LongestOption's lowest Brier score despite low accuracy confirms that probability quality and discrimination answer different questions.

Qwen's raw valid-label confidence was poorly calibrated. Isotonic regression reduced direct-inference ECE from 0.2646 to 0.0147 and self-verification ECE from 0.5805 to 0.0340. It also reduced both Brier score and NLL, while the decoded answers remained fixed.

Table 11. Qwen confidence calibration before and after isotonic regression.

Condition	Mapping	ECE-10	Brier	NLL	AURC
Direct	Raw	0.2646	0.3281	1.7119	0.6384
Direct	Isotonic	0.0147	0.2190	0.6447	0.6217
Self-verification	Raw	0.5805	0.5691	2.6239	0.6662
Self-verification	Isotonic	0.0340	0.2062	0.6099	0.6611

Figure 4. Qwen direct reliability before and after validation-fitted isotonic regression.



The calibrated direct scores clustered at a few values because the 512-item sample supplied limited resolution. Reliability improved, but AURC remained 0.6217, indicating weak difficulty ordering.

This separation between calibration and ranking is consistent with cross-domain reports in credit risk, privacy-constrained policy estimation, and sensor fusion: a post-hoc mapping can improve the meaning of a score without changing the underlying prediction or guaranteeing useful ordering under shift (Bai, Wang, et al., 2026; Jin et al., 2024b; Xin, 2025b). AgriEval exhibits the same distinction more sharply because the calibrated Qwen scores became reliable in aggregate while the highest-confidence subset still contained many errors.

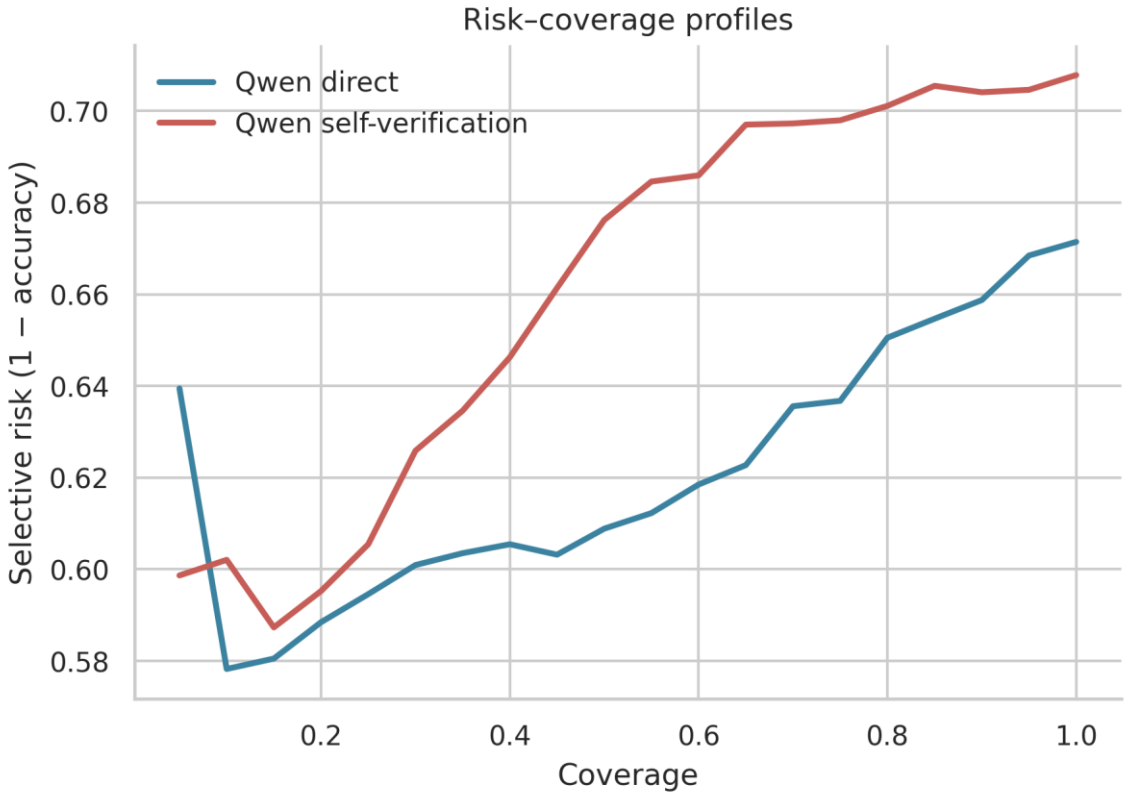
4.4 Selective prediction, self-verification, and routing

Confidence ranking did not yield a low-risk operating region. Direct risk was 0.5782 at 10% coverage and 0.6088 at 50%; self-verification risk was 0.6020 and 0.6762 at the same coverages. Both approached their full-coverage error rates as more cases were retained.

Table 12. Qwen calibrated selective risk by retained coverage.

Coverage	Direct retained n	Direct risk	Self-verification retained n	Self-verification risk
0.10	294	0.5782	294	0.6020
0.25	735	0.5946	735	0.6054
0.50	1,470	0.6088	1,470	0.6762
0.75	2,205	0.6367	2,205	0.6980
1.00	2,940	0.6714	2,940	0.7078

Figure 5. Risk–coverage curves for Qwen direct and self-verification after calibration.



Selective classification reduces risk only when confidence ranks cases effectively (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017). Here, the most confident tenth still erred above 0.57.

The second pass changed 43.30% of predictions. It corrected 261 initial errors but changed 368 initially correct answers to wrong ones. Accuracy consequently declined by 3.64 points, and the exact McNemar test rejected equal paired error rates ((p<.001)).

Table 13. Direct-to-self-verification outcome transitions.

Direct outcome	Self-verification outcome	n	Share of all cases (%)
Correct	Correct	598	20.34
Correct	Wrong	368	12.52
Wrong	Correct	261	8.88
Wrong	Wrong	1,713	58.27

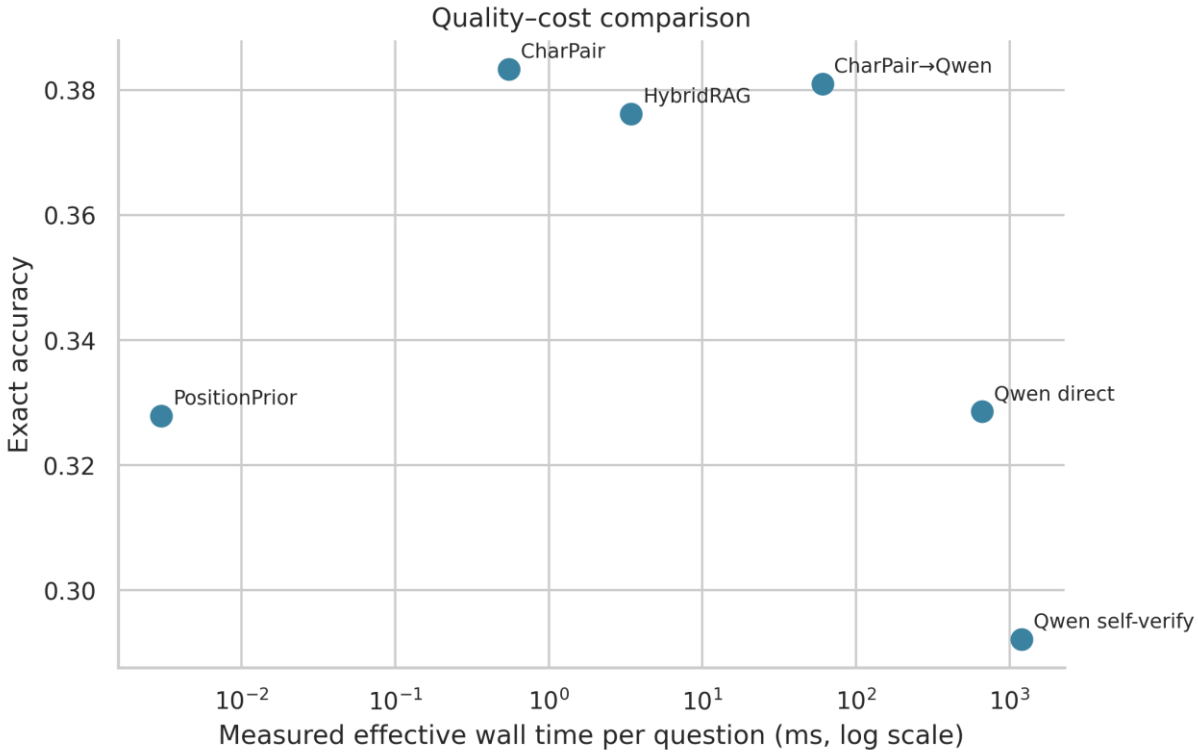
The pass corrected 13.22% of direct errors but corrupted 38.10% of direct successes. It added 534.122 ms per question and reduced accuracy.

The validation-selected router escalated cases with calibrated CharPair confidence below 0.1951. It routed 9.01% of test questions to Qwen, reached 0.3810 accuracy, and cost 60.455 ms per question. CharPair alone was both more accurate, at 0.3833, and about 110 times faster.

Table 14. Calibration-trained router compared with its component systems.

System	Qwen route rate	Exact accuracy	Option micro-F1	ECE-10	Effective ms/question	Effective ms/correct
CharPair	0	0.3833	0.5580	0.0347	0.5491	1.4324
Qwen direct	1.0000	0.3286	0.5099	0.0147	664.6151	2,022.7415
CharPair→Qwen router	0.0901	0.3810	0.5563	0.0341	60.4549	158.6940

Figure 6. Measured multiple-choice accuracy–compute frontier, including the validation-selected router.

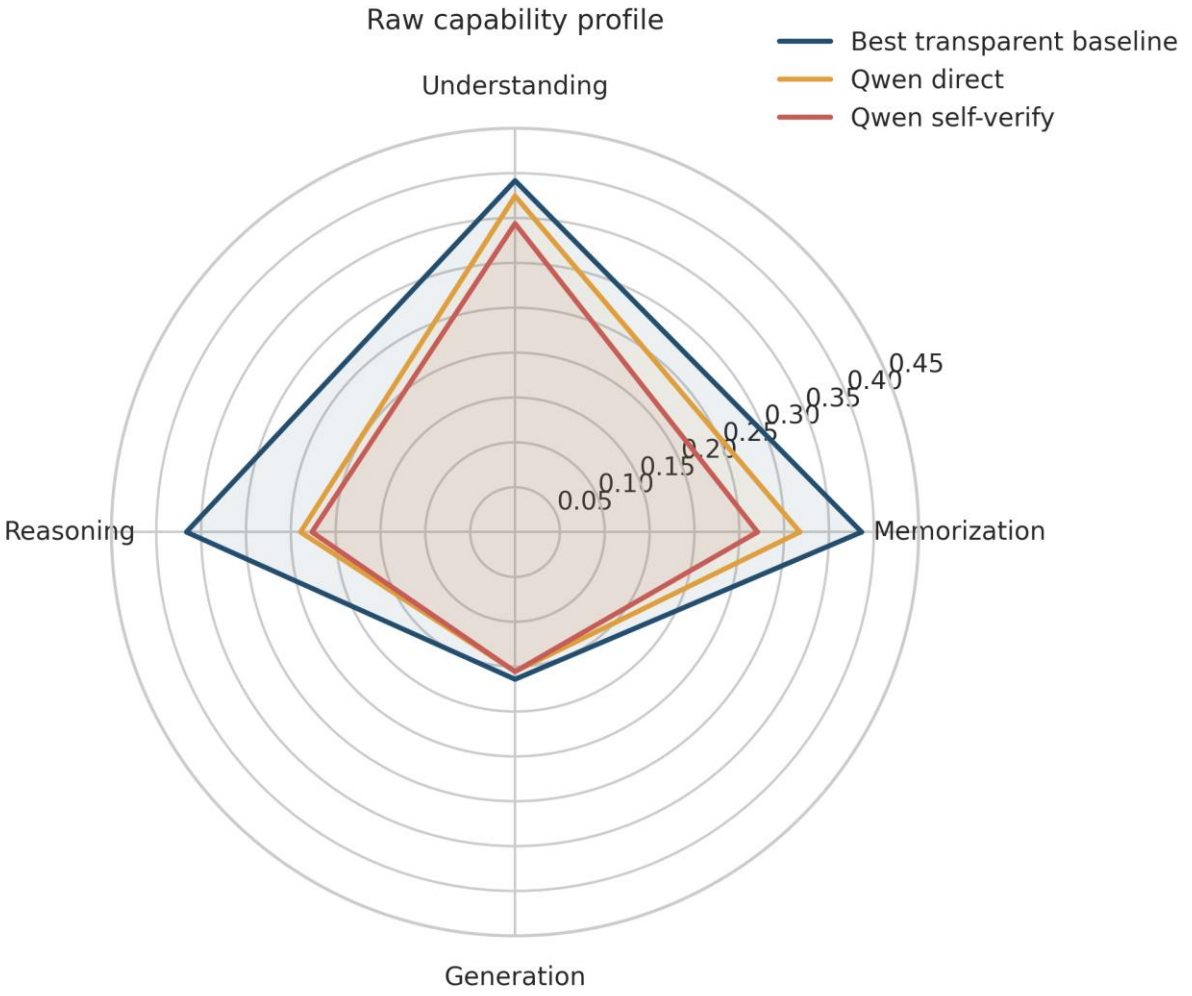


Few escalations kept router accuracy near CharPair, but Qwen supplied no quality gain. Routing therefore requires held-out accuracy and measured cost.

This negative result is operationally informative. Cost-aware routing in AIOps and resource admission is valuable only when the escalated path has a demonstrable conditional advantage, and offline policy evaluation warns that a policy can be misranked when validation support is weak (C. Li et al., 2026; Mu et al., 2025; Zhao et al., 2026). Here the validation-selected route rate was small, but the expensive path was not better on the routed test cases; calibration alone could not create that missing complementarity.

Figure 7 reports each system's raw score on its corresponding axis. Memorization, understanding, and reasoning are exact multiple-choice accuracies; generation is character ROUGE-L. The generation scale is not commensurate with the three accuracy axes.

Figure 7. Raw capability profile for the best transparent baseline bundle, Qwen direct, and Qwen self-verification.



4.5 Open-answer retrieval and generation

VerifiedRAG led open answers with ROUGE-L 0.1641, chrF 0.0963, cosine 0.0835, and ROUGE-L ≥ 0.25 rate 0.1917. DomainRAG was close and exceeded unrestricted CharRAG by 1.47 ROUGE-L points, supporting domain filtering. Qwen generation obtained ROUGE-L 0.1555, chrF 0.0815, cosine 0.0745, and an overlap-success rate of 0.1155.

Table 15. Open-answer test performance.

System	ROUGE-L	chrF	TF-IDF cosine	ROUGE-L ≥ 0.25 rate	Scoring latency (ms/q)
CharRAG	0.1485	0.0864	0.0735	0.1501	0.0038
DomainRAG	0.1633	0.0957	0.0834	0.1871	0.0138
HybridRAG	0.1598	0.0932	0.0803	0.1801	0.0226
VerifiedRAG	0.1641	0.0963	0.0835	0.1917	0.0326
Qwen generation	0.1555	0.0815	0.0745	0.1155	912.1602

The Qwen run generated a mean of 34.49 tokens, and 12.24% of responses reached the 64-token limit. Its ROUGE-L difference from VerifiedRAG was −0.0085 with a paired 95% interval of [−0.0215, 0.0041], so these data did not establish a reliable difference in reference overlap. The metrics measure overlap rather than semantic or agronomic validity: repeated crop or treatment terms can score well despite an unsuitable recommendation.

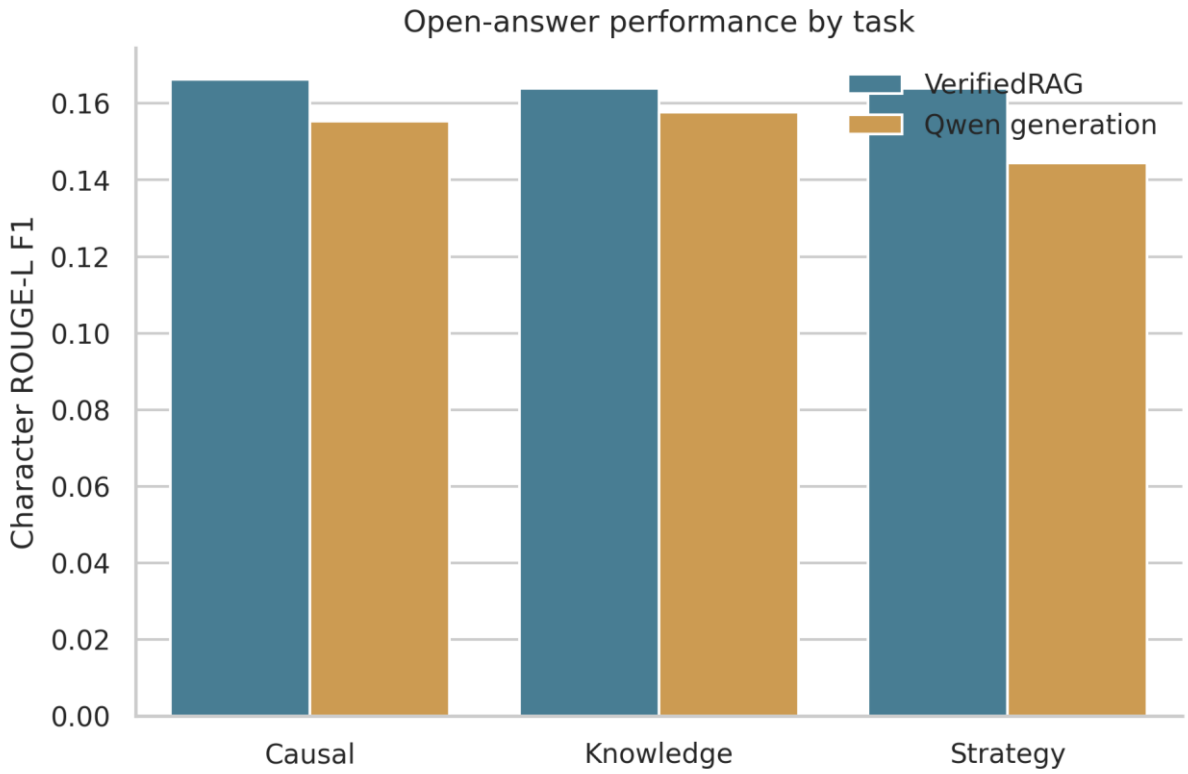
Evidence-oriented RAG work suggests the next evaluation layer: retrieve an authoritative source, measure evidence coverage, and test whether each generated claim is supported before presenting it to an expert (Jin, 2025a; C. Li et al., 2024; Su, Chen, et al., 2026; Zhong, Chen, et al., 2025). The present answer-copy systems do not meet that standard because the training answers are cases rather than an independently curated evidence corpus.

VerifiedRAG ROUGE-L was 0.1662/0.1639/0.1639 for causal, knowledge, and strategy tasks. Qwen reached 0.1553/0.1577/0.1444. DomainRAG narrowly led the 28-item causal subset; VerifiedRAG led the larger knowledge and strategy subsets.

Table 16. Open-answer performance by generation task.

System	Task	n	ROUGE-L	chrF	TF-IDF cosine
CharRAG	Causal	28	0.1423	0.0965	0.0797
CharRAG	Knowledge	340	0.1498	0.0861	0.0682
CharRAG	Strategy	65	0.1446	0.0839	0.0983
DomainRAG	Causal	28	0.1679	0.1116	0.0907
DomainRAG	Knowledge	340	0.1633	0.0941	0.0768
DomainRAG	Strategy	65	0.1613	0.0974	0.1153
HybridRAG	Causal	28	0.1617	0.1086	0.0898
HybridRAG	Knowledge	340	0.1600	0.0917	0.0742
HybridRAG	Strategy	65	0.1578	0.0942	0.1080
VerifiedRAG	Causal	28	0.1662	0.1100	0.0897
VerifiedRAG	Knowledge	340	0.1639	0.0948	0.0772
VerifiedRAG	Strategy	65	0.1639	0.0981	0.1136
Qwen generation	Causal	28	0.1553	0.0777	0.0746
Qwen generation	Knowledge	340	0.1577	0.0813	0.0738
Qwen generation	Strategy	65	0.1444	0.0845	0.0780

Figure 8. ROUGE-L by open-answer task for VerifiedRAG and Qwen generation.



Answer copying establishes a lexical retrieval floor, not generative capability. Qwen produced new text, but its low absolute overlap and frequent length-limit termination show that a 0.49-billion-parameter model with greedy decoding did not reliably reconstruct the reference content.

4.6 Statistical uncertainty and integrated capability profile

Bootstrap intervals were [0.3653, 0.3997] for CharPair accuracy, [0.3119, 0.3459] for Qwen direct accuracy, and [0.1477, 0.1636] for Qwen ROUGE-L. They cover test-item sampling, not alternative splits, annotations, prompts, or hardware.

Table 17. Bootstrap intervals and paired comparisons.

Comparison	Estimate	95% CI	Procedure	p
CharPair multiple-choice exact accuracy	0.3833	[0.3653, 0.3997]	Item bootstrap, 1,000 replicates	—
Qwen direct exact accuracy	0.3286	[0.3119, 0.3459]	Item bootstrap, 5,000 replicates	—
Qwen self-verification exact accuracy	0.2922	[0.2765, 0.3088]	Item bootstrap, 5,000 replicates	—
VerifiedRAG open-answer ROUGE-L	0.1641	[0.1524, 0.1761]	Item bootstrap, 1,000 replicates	—
Qwen generation ROUGE-L	0.1555	[0.1477, 0.1636]	Item bootstrap, 5,000 replicates	—
Qwen direct − CharPair accuracy	−0.0548	[−0.0769, −0.0316]	Paired bootstrap; exact McNemar	<.001
Qwen self-verification − direct accuracy	−0.0364	[−0.0531, −0.0197]	Paired bootstrap; exact McNemar	<.001
Qwen generation − VerifiedRAG ROUGE-L	−0.0085	[−0.0215, 0.0041]	Paired bootstrap, 5,000 replicates	—

The multiple-choice intervals exclude zero for both paired contrasts: CharPair exceeded Qwen direct, and self-verification harmed Qwen. The open-answer interval crosses zero, so its smaller observed difference warrants a more cautious interpretation. Overall, character discrimination beat case retrieval; fusion helped reasoning; isotonic regression improved confidence interpretation without improving answers; and neither the critique pass nor the selected routing policy justified its extra computation.

Plant-production dominance, small subgroups, Chinese-language scope, one quantized small model, a single deterministic prompt, binary exact-correctness calibration, and overlap-only open metrics limit generalization. Latencies are measured on one CPU environment and should be interpreted as reproducible local costs rather than hardware-independent constants.

5. Conclusions

This study evaluated Qwen2.5-0.5B-Instruct and transparent reference systems on 16,864 AgriEval questions after an integrity audit and group-aware split. The corpus covered six major agricultural domains, 29 subdomains, and 15 tasks, but plant production accounted for 76.08% of records. Grouping normalized stems before splitting kept repeated questions within one fold and made the held-out comparisons less vulnerable to duplicate leakage.

CharPair was the strongest pooled multiple-choice system, with 0.3833 exact accuracy and a 95% interval of [0.3653, 0.3997]. Qwen direct reached 0.3286, while HybridRAG led the reasoning subset at 0.3664. The contrast shows why pooled results and capability profiles must be read together. Qwen's 0.2865 exact-set accuracy but 0.8362 option micro-F1 on multiple selection further shows that partial label recovery can coexist with frequent set-level failure.

Isotonic regression reduced Qwen direct ECE from 0.2646 to 0.0147, but selective risk remained high. Self-verification reduced accuracy from 0.3286 to 0.2922 because 368 correct initial answers were corrupted while only 261 errors were corrected. The calibration-trained router likewise failed to improve on CharPair: it routed 9.01% of cases, reached 0.3810 accuracy, and increased cost from 0.549 to 60.455 ms per question.

For open answers, VerifiedRAG reached ROUGE-L 0.1641 and Qwen reached 0.1555; their paired interval included zero. These overlap metrics do not establish agronomic correctness. Human expert review of factuality, safety, and actionable completeness is therefore the necessary next step for generation, alongside tests of larger models, alternative prompts, and evidence-grounded retrieval.

The practical conclusion is that agricultural language evaluation is a profile rather than one score. A useful system must beat simple lexical and positional references, remain stable in weak domains, express uncertainty honestly, and justify additional computation with a measured quality gain.

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] Bai, J., Chen, S., Zheng, D., & Kuo, M.-J. (2026). Interpretable attack-chain stage detection from AWS CloudTrail event sequences via linear models and HMM smoothing. *Information Electrical and Electronic Engineering, 6*(1), 28–43. https://doi.org/10.33474/infotron.v6i1.24923

[2] Bai, J., Wang, H., Wu, Q., & Zhang, B. (2026). Privacy-robust incrementality estimation in cookieless settings via uplift modeling: Reproducible evidence from the Hillstrom e-mail experiment. *Journal of Technology Informatics and Engineering, 5*(1), 17–38. https://doi.org/10.51903/jtie.v5i1.468

[3] Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. David McKay Company.

[4] Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In Y. Lechevallier & G. Saporta (Eds.), *Proceedings of COMPSTAT'2010* (pp. 177–186). Physica-Verlag. https://doi.org/10.1007/978-3-7908-2604-3_16

[5] Chang, X., Lu, Y., & Zhong, Z. S. (2026). Review-grounded explainable recommendation with faithfulness evaluation on Amazon reviews. *Journal of Electrical Engineering and Computer Science, 11*(1), 9–22. https://doi.org/10.54732/jeecs.v11i1.2

[6] 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/forum?id=cSimKw5p6R

[7] Chen, S., He, S., & Sun, E. (2024). Risk-bounded GPU resource oversubscription via conformal demand envelopes in production AI clusters. *Journal of Advanced Computing Systems, 4*(5), 119–134. https://doi.org/10.69987/jacs.2024.40509

[8] Chen, Y., & Xu, H. (2026). Trust-calibrated multilingual RAG for humanitarian information platforms: Empirical evaluation on OMoS-QA for migration information access. *International Journal of Graphic Design, 4*(1), 141–164. https://doi.org/10.51903/ijgd.v4i1.3552

[9] Chen, Y., Zhang, Y., Chau, D., & Sherman, M. (2023). Credit card default risk tiering with probability calibration and uncertainty-driven rejection: A reproducible study on the UCI credit card clients dataset. *Journal of Advanced Computing Systems, 3*(4), 31–47. https://doi.org/10.69987/jacs.2023.30403

[10] Chen, Y., Zhang, Y., & Sherman, M. (2024). Going concern and bankruptcy prediction under extreme class imbalance: Cost-sensitive learning, resampling, and focal loss with explainable financial-ratio portraits. *Journal of Advanced Computing Systems, 4*(4), 80–96. https://doi.org/10.69987/jacs.2024.40407

[11] Chen, Y., Zhou, S., & Lin, E. (2025). Accounting-aware evidence retrieval for institutional due diligence of tokenized trade receivable RWA. *Journal of Technology Informatics and Engineering, 4*(3), 649–663. https://doi.org/10.51903/jtie.v4i3.542

[12] Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory, 13*(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964

[13] De Clercq, D., Nehring, E., Mayne, H., & Mahdi, A. (2024). Large language models can help boost food production, but be mindful of their risks. *Frontiers in Artificial Intelligence, 7*, Article 1326153. https://doi.org/10.3389/frai.2024.1326153

[14] Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics, 7*(1), 1–26. https://doi.org/10.1214/aos/1176344552

[15] El-Yaniv, R., & Wiener, Y. (2010). On the foundations of noise-free selective classification. *Journal of Machine Learning Research, 11*, 1605–1641. https://www.jmlr.org/papers/v11/el-yaniv10a.html

[16] Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. *Advances in Neural Information Processing Systems, 30*, 4878–4887. https://proceedings.neurips.cc/paper/7073-selective-classification-for-deep-neural-networks

[17] 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

[18] Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M.-W. (2020). REALM: Retrieval-augmented language model pre-training. *Proceedings of Machine Learning Research, 119*, 3929–3938. https://proceedings.mlr.press/v119/guu20a.html

[19] He, S., Chang, X., & Sun, E. (2024). Cross-cloud transfer learning for AI training capacity forecasting under workload and topology distribution shift. *Journal of Advanced Computing Systems, 4*(1), 100–120. https://doi.org/10.69987/jacs.2024.40108

[20] He, S., Li, C., & Rao, H. (2025). Few-shot cold-start workload forecasting for new AI inference tenants with time-series foundation models. *Journal of Technology Informatics and Engineering, 4*(1), 306–324. https://doi.org/10.51903/jtie.v4i1.546

[21] He, S., Nie, J., & Li, C. (2026). Power-aware inventory planning for AI infrastructure using job-level forecasting and LLM workload explanations. *Journal of Technology Informatics and Engineering, 5*(1), 341–359. https://doi.org/10.51903/jtie.v5i1.548

[22] 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

[23] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring massive multitask language understanding. In *International Conference on Learning Representations*. https://openreview.net/forum?id=d7KBjmI3GmQ

[24] Huang, Y., Bai, Y., Zhu, Z., Zhang, J., Zhang, J., Su, T., Liu, J., Lv, C., Zhang, Y., Lei, J., Fu, Y., Sun, M., & He, J. (2023). C-Eval: A multi-level multi-discipline Chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems, 36*, 62991–63010. https://doi.org/10.52202/075280-2749

[25] Jin, J. (2025a). Evidence-chain reliable RAG: Hallucination detection, source attribution, and deterministic provenance explanations. *Journal of Technology Informatics and Engineering, 4*(2), 520–533. https://doi.org/10.51903/jtie.v4i2.535

[26] Jin, J. (2025b). LLM-style evidence cards for scientific search interfaces: A UI/UX design framework for retrieval transparency, ranking trust, and visual evidence hierarchy. *International Journal of Graphic Design, 3*(2), 397–414. https://doi.org/10.51903/ijgd.v3i2.3698

[27] Jin, J., Huang, T., & Lu, S. (2024a). Cost-sensitive learning, simulated PU learning, and one-class autoencoding for extreme-imbalance credit card fraud detection. *Journal of Advanced Computing Systems, 4*(6), 64–73. https://doi.org/10.69987/jacs.2024.40605

[28] Jin, J., Huang, T., & Lu, S. (2024b). A model-risk-friendly probability of default workflow: Calibration, distribution-free uncertainty quantification, and SHAP explanations on the UCI credit card default dataset. *Journal of Advanced Computing Systems, 4*(6), 74–85. https://doi.org/10.69987/jacs.2024.40606

[29] Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-T. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 6769–6781). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.550

[30] Kuo, M.-J., Zheng, D., & Hires, J. (2025). Federated topic-preference learning for knowledge-grounded chat with differential privacy. *Journal of Technology Informatics and Engineering, 4*(2), 385–401. https://doi.org/10.51903/jtie.v4i2.502

[31] 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://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[32] Li, C., Bai, J., & Wang, S. (2024). Evidence-chain reliable RAG: Word-level hallucination detection, source attribution, and provenance explanation for LLM applications. *Journal of Advanced Computing Systems, 4*(2), 76–92. https://doi.org/10.69987/jacs.2024.40207

[33] Li, C., Liu, G., & Zhao, Z. (2026). Cost-aware LLM-style routing for AIOps log analysis: Log parsing, anomaly detection, fault diagnosis, and incident summarization on LogEval task files. *Journal of Technology Informatics and Engineering, 5*(2), 91–103. https://doi.org/10.51903/jtie.v5i2.538

[34] Li, C., Zhou, B., & Gao, K. (2025). Risk-calibrated patient-facing AI safety cards: A UI/UX benchmark for explainable medical AI response interfaces. *International Journal of Graphic Design, 3*(2), 381–394. https://doi.org/10.51903/ijgd.v3i2.3709

[35] Li, H., Wu, H., Li, Q., & Zhao, C. (2025). A review on enhancing agricultural intelligence with large language models. *Artificial Intelligence in Agriculture, 15*(4), 671–685. https://doi.org/10.1016/j.aiia.2025.05.006

[36] Li, H., Zhang, Y., Koto, F., Yang, Y., Zhao, H., Gong, Y., Duan, N., & Baldwin, T. (2024). CMMLU: Measuring massive multitask language understanding in Chinese. In *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 11260–11285). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.671

[37] Li, J., & Zhou, A. (2026). Multi-regulation RAG for AI product counsel: A legal governance framework for cross-border digital commerces. *Rule of Law Studies Journal, 2*(2), 105–123. https://doi.org/10.64780/rolsj.v2i2.225

[38] Li, Y. (2024). Findable then explainable: Retrieval–summary integration for code intelligence on a lightweight CodeSearchNet subset. *Journal of Advanced Computing Systems, 4*(7), 65–82. https://doi.org/10.69987/jacs.2024.40706

[39] Li, Y., & Lu, S. (2025). Language-guided feature selection for DDoS and intrusion detection on CICIDS2017. *Journal of Technology Informatics and Engineering, 4*(1), 284–305. https://doi.org/10.51903/jtie.v4i1.531

[40] Li, Z., Zhang, K., & Wong, A. (2026). Numerical-reasoning guardrails for a quant research assistant: A compact reproducible benchmark using SEC and FRED data. *Journal of Technology Informatics and Engineering, 5*(2), 75–90. https://doi.org/10.51903/jtie.v5i2.541

[41] Li, Z., Zhou, S., & Zhou, Z. (2025). Financial risk dashboard design for institutional RWA investors: Visual hierarchy, chart comprehension, and explainability in FinChart-Bench. *International Journal of Graphic Design, 3*(1), 196–210. https://doi.org/10.51903/ijgd.v3i1.3715

[42] Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out* (pp. 74–81). Association for Computational Linguistics. https://aclanthology.org/W04-1013/

[43] Liu, G., He, S., & Liu, I. (2023). LLM-augmented multi-source root cause attribution for CPU and network faults in microservices. *Journal of Advanced Computing Systems, 3*(6), 39–57. https://doi.org/10.69987/jacs.2023.30604

[44] Liu, G., He, S., & Wong, H. (2025). LLM-compatible visual brief cards for AI infrastructure capacity dashboards: A UI/UX framework for turning forecast risk into graphic design decisions. *International Journal of Graphic Design, 3*(1), 196–213. https://doi.org/10.51903/ijgd.v3i1.3723

[45] Liu, G., Li, C., & Zhang, E. (2024). OpsLLM for cloud incident triage: Bilingual RAG-based root cause analysis and alert summarization for AI infrastructure operations. *Journal of Advanced Computing Systems, 4*(4), 97–111. https://doi.org/10.69987/jacs.2024.40408

[46] Lu, S., & Zhou, D. (2024). TinyLLM-assisted intrusion detection for real-time IoT networks. *Journal of Advanced Computing Systems, 4*(8), 72–87. https://doi.org/10.69987/jacs.2024.40809

[47] McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika, 12*(2), 153–157. https://doi.org/10.1007/BF02295996

[48] Meng, S., Chen, J., & Zheng, I. (2026). LLM-inspired offline reranking for financial search: Query rewriting, hybrid retrieval, and listwise relevance ranking on FiQA. *Journal of Technology Informatics and Engineering, 5*(1), 361–378. https://doi.org/10.51903/jtie.v5i1.537

[49] Mi, G., Ye, T., & Wood, D. (2025). A lightweight medical foundation model for cross-modal multi-task pretraining and parameter-efficient few-shot transfer on MedMNIST. *Journal of Technology Informatics and Engineering, 4*(3), 572–589. https://doi.org/10.51903/jtie.v4i3.492

[50] Mu, J., Lu, Y., & Smith, M. (2023). LLM-assisted incrementality (uplift) modeling for email advertising: From feature interactions to interpretable audience–creative–channel policies. *Journal of Advanced Computing Systems, 3*(1), 31–48. https://doi.org/10.69987/jacs.2023.30103

[51] Mu, J., Ye, T., & Patel, P. (2025). Offline counterfactual evaluation for advertising and recommendation slot policies: A reproducible study on the Open Bandit Dataset (small). *Journal of Technology Informatics and Engineering, 4*(3), 521–543. https://doi.org/10.51903/jtie.v4i3.500

[52] Nie, J., Liu, G., Li, C., & Zou, T. (2026). Evidence-constrained incident visualization cards for distributed cloud logs: A UI/UX framework for turning Hadoop, OpenStack, and ZooKeeper logs into actionable SRE design interfaces. *International Journal of Graphic Design, 4*(1), 179–185. https://doi.org/10.51903/ijgd.v4i1.3703

[53] Nie, J., & Zheng, D. (2024). Noisy-neighbor-aware VM degradation risk modeling with unsupervised residual fusion. *Journal of Advanced Computing Systems, 4*(4), 112–123. https://doi.org/10.69987/jacs.2024.40409

[54] Ong, I., Almahairi, A., Wu, V., Chiang, W.-L., Wu, T., Gonzalez, J. E., Kadous, M. W., & Stoica, I. (2025). RouteLLM: Learning to route LLMs from preference data. In *The Thirteenth International Conference on Learning Representations*. https://proceedings.iclr.cc/paper_files/paper/2025/hash/5503a7c69d48a2f86fc00b3dc09de686-Abstract-Conference.html

[55] 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*, 2825–2830. https://www.jmlr.org/papers/v12/pedregosa11a.html

[56] Pezeshkpour, P., & Hruschka, E. (2024). Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024* (pp. 2006–2017). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-naacl.130

[57] Platt, J. C. (2000). Probabilities for SV machines. In A. J. Smola, P. L. Bartlett, B. Schölkopf, & D. Schuurmans (Eds.), *Advances in large-margin classifiers* (pp. 61–74). MIT Press. https://doi.org/10.7551/mitpress/1113.003.0008

[58] Popović, M. (2015). chrF: Character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation* (pp. 392–395). Association for Computational Linguistics. https://doi.org/10.18653/v1/W15-3049

[59] Qwen Team. (2025). *Qwen2.5 technical report*. arXiv. https://doi.org/10.48550/arXiv.2412.15115

[60] Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management, 24*(5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0

[61] Shaikh, T. A., Rasool, T., Veningston, K., & Yaseen, S. M. (2025). The role of large language models in agriculture: Harvesting the future with LLM intelligence. *Progress in Artificial Intelligence, 14*, 117–164. https://doi.org/10.1007/s13748-024-00359-4

[62] Su, W., Chen, S., & Qian, E. (2026). Narrative-aware scientific claim verification agent with evidence ranking for ClimateCheck. *Journal of Technology Informatics and Engineering, 5*(1), 327–340. https://doi.org/10.51903/jtie.v5i1.549

[63] Su, W., Chen, S., & Zhao, C. (2025). Budgeted multi-hop retrieval agent for compositional question answering: A retrieval-policy evaluation on the official MultiHop-RAG benchmark. *Journal of Technology Informatics and Engineering, 4*(3), 649–662. https://doi.org/10.51903/jtie.v4i3.543

[64] Su, W., Rao, H., & Ma, E. (2026). Privacy and data-integrity risk cards for LLM agents: A UI/UX design framework for secure human oversight under prompt-injection attacks. *International Journal of Graphic Design, 4*(1), 186–191. https://doi.org/10.51903/ijgd.v4i1.3699

[65] Sun, X., Lu, Y., & Chen, J. (2023). Controllable long-term user memory for multi-session dialogue: Confidence-gated writing, time-aware retrieval-augmented generation, and update/forgetting. *Journal of Advanced Computing Systems, 3*(8), 9–24. https://doi.org/10.69987/jacs.2023.30802

[66] Sun, X., Zhong, Z. S., & Wu, Q. (2026). Retrieval-grounded HDFS log anomaly detection and deterministic failure narrative generation. *Journal of Computational Systems and Applications, 3*(1), 15–30. https://doi.org/10.64229/j6d7fr94

[67] Tzachor, A., Devare, M., Richards, C., Pypers, P., Ghosh, A., Koo, J., Johal, S., & King, B. (2023). Large language models and agricultural extension services. *Nature Food, 4*(11), 941–948. https://doi.org/10.1038/s43016-023-00867-x

[68] Wu, Q., Mi, G., & Wood, D. (2025). Calibration-light subject-independent motor imagery BCI via self-supervised pretraining and conformer. *Journal of Technology Informatics and Engineering, 4*(1), 239–262. https://doi.org/10.51903/jtie.v5i1.493

[69] Xin, Q. (2025a). Explaining OpenStack failure-injection log anomalies with retrieved normal prototypes. *Emerging Information Science and Technology, 6*(2), 125–146. https://doi.org/10.18196/eist.v6i2.31232

[70] Xin, Q. (2025b). Uncertainty-aware late fusion for 3D perception (confidence calibration + fusion rule learning). *Journal of Technology Informatics and Engineering, 4*(1), 215–238. https://doi.org/10.51903/jtie.v4i1.485

[71] Xin, Q. (2026a). Auditable automated essay scoring and formative feedback: A rubric-grounded pipeline for secondary and higher education. *Journal of Applied Artificial Intelligence in Education, 2*(1), 1–19. https://doi.org/10.66053/jaaie.v2i1.348

[72] Xin, Q. (2026b). Early-warning analytics with LLM intervention rationales for student retention decisions: Classroom interaction modeling with xAPI-edu-data and dropout/success prediction. *Interdisciplinary Journal of Pedagogy and Research in Media Technology, 2*(1), 9–27. https://doi.org/10.64268/inspire.v2i1.117

[73] Xin, Q. (2026c). Explainable and fair credit risk scoring with counterfactual explanations: A reproducible evaluation on the German Credit dataset (HELOC-motivated). *J-INTECH (Journal of Information and Technology), 14*(2), 215–231. https://doi.org/10.32664/j-intech.v14i02.2228

[74] Xin, Q. (2026d). Host-based intrusion detection with system call sequences: Window localization and forensic narratives. *Aviation Electronics, Information Technology, Telecommunications, Electricals, and Controls, 8*(2), 325–334. https://doi.org/10.28989/avitec.v8i2.3973

[75] Xin, Q. (2026e). LiDAR–camera object-level fusion for multi-target tracking using JPDA and EKF: A reproducible empirical study on a PandaSet-parameterised five-sequence dataset. *Journal of Technology Informatics and Engineering, 5*(1), 54–76. https://doi.org/10.51903/jtie.v5i1.486

[76] Xin, Q. (2026f). Log anomaly detection with conformal alert control and evidence-grounded incident ticket generation. *Aviation Electronics, Information Technology, Telecommunications, Electricals, and Controls, 8*(2), 247–264. https://doi.org/10.28989/avitec.v8i2.3974

[77] Xin, Q. (2026g). Probabilistic bike-sharing demand forecasting under changing weather and seasonal regimes with transformer-based models. *Findings*. https://doi.org/10.32866/001c.157499

[78] Xin, Q. (2026h). Self-supervised customer representation learning for segmentation and next-purchase prediction on UCI online retail. *J-INTECH (Journal of Information and Technology), 14*(1), 20–37. https://doi.org/10.32664/j-intech.v14i01.2229

[79] Xin, Q. (2026i). Self-supervised log anomaly detection with LogBERT-style transformers: Full empirical evaluation on a reproducible SynHDFS benchmark. *Journal of Electrical Engineering and Computer Science, 11*(1), 23–35. https://doi.org/10.54732/jeecs.v11i1.3

[80] Yan, L., Wang, H., Tang, C., Liu, H., Sun, T., Liu, L., Guan, Y., & Jiang, J. (2026). AgriEval: A comprehensive Chinese agricultural benchmark for large language models. *Proceedings of the AAAI Conference on Artificial Intelligence, 40*(40), 34205–34213. https://doi.org/10.1609/aaai.v40i40.40716

[81] Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 694–699). Association for Computing Machinery. https://doi.org/10.1145/775047.775151

[82] Zhang, B., Rao, H., & Zhao, D. (2024). Evidence-grounded RAG for cloud-native DevOps: Hallucination-resistant AIOps question answering over private operations documents. *Journal of Advanced Computing Systems, 4*(3), 109–125. https://doi.org/10.69987/jacs.2024.40308

[83] Zhang, B., Ren, Y., & Zou, J. (2025). LLM-style explainable e-commerce recommendation cards: A UI/UX design framework for trust-calibrated product recommendation. *International Journal of Graphic Design, 3*(2), 381–396. https://doi.org/10.51903/ijgd.v3i2.3697

[84] Zhang, B., Sun, X., Liu, G., & Zhou, B. (2026). LLM-style DevOps copilot for cloud-native troubleshooting: Retrieval-augmented runbook generation and command-safety evaluation. *Journal of Technology Informatics and Engineering, 5*(2), 104–118. https://doi.org/10.51903/jtie.v5i2.534

[85] Zhang, J. (2025). From general human activity recognition to volleyball-oriented wearable transfer learning: Cross-dataset evidence from UCI HAR and WISDM for domain adaptation and edge deployment. *Journal of Technology Informatics and Engineering, 4*(1), 263–283. https://doi.org/10.51903/jtie.v4i1.524

[86] Zhang, J. (2026). Early warning, grade prediction, and teacher-facing LLM-ready explanations toward an open volleyball course: Reproducible evidence from four public education datasets. *Journal of Technology Informatics and Engineering, 5*(2), 20–44. https://doi.org/10.51903/jtie.v5i2.525

[87] Zhang, K., Chen, Y., & Qian, A. (2025). Evidence-grounded accounting disclosure review cards: A visual communication framework for LLM-style explanations over SEC financial statements and notes. *International Journal of Graphic Design, 3*(2), 395. https://doi.org/10.51903/ijgd.v3i2.3710

[88] Zhang, R., Wen, Z., Wang, C., Tang, C., Xu, P., & Jiang, Y. (2025). *Quality analysis and evaluation prediction of RAG retrieval based on machine learning algorithms* [Preprint]. arXiv. https://doi.org/10.48550/arxiv.2511.19481

[89] Zhang, Y., & Zhang, H. (2025a). A therapist-facing session copilot for live counseling support: Reasoning-guided retrieval and ranking from multi-turn counseling dialogues. *Journal of Technology Informatics and Engineering, 4*(2), 464–486. https://doi.org/10.51903/jtie.v4i2.547

[90] Zhang, Y., & Zhang, H. (2025b). Visualizing the right counseling support: Evidence-linked recommendation cards for explainable mental health intake interfaces. *International Journal of Graphic Design, 3*(1), 214–229. https://doi.org/10.51903/ijgd.v3i1.3722
[91] Zhang, Y., & Zhou, Z. (2026). Strategy-aware therapist imitation for emotional support dialogues: A reproducible ESConv study for LLM response control. *Advances in Educational Technology and Psychology, 10*(2), 92–97. https://doi.org/10.23977/aetp.2026.100213
[92] Zhao, S., Bai, J., & Roberson, D. (2025). Multi-horizon GPU demand forecasting with workload semantics and operational risk curves: An empirical study on Alibaba clusterdata GPU trace. *Journal of Technology Informatics and Engineering, 4*(3), 544–571. https://doi.org/10.51903/jtie.v4i3.498
[93] Zhao, S., Ren, Y., & Chang, X. (2026). Profit-aware spot GPU admission control with cost-sensitive loss and evidence-grounded policy memos for AI workload supply-demand matching. *Journal of Technology Informatics and Engineering, 5*(2), 45–59. https://doi.org/10.51903/jtie.v5i2.545
[94] Zheng, C., Zhou, H., Meng, F., Zhou, J., & Huang, M. (2024). Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=shr9PXz7T0
[95] Zheng, D., & Li, C. (2024). Behavior-level jailbreak resistance via multi-stage refusal + utility preservation. *Journal of Advanced Computing Systems, 4*(1), 83–99. https://doi.org/10.69987/jacs.2024.40107
[96] Zheng, D., Li, C., & Davidson, H. (2023). Continual red-teaming for in-the-wild jailbreaks via online guardrail updates and guardrail distillation. *Journal of Advanced Computing Systems, 3*(2), 35–49. https://doi.org/10.69987/jacs.2023.30203
[97] Zheng, D., Zhang, B., & Geibel, J. (2024). VerifySafe: Toxicity-safe agent responses under adversarial prompts with evidence-based self-verification. *Journal of Advanced Computing Systems, 4*(1), 67–82. https://doi.org/10.69987/jacs.2024.40106
[98] Zhong, Z. S., Chen, J., Zhong, E., & Sun, X. (2025). Evidence-calibrated RAG for unanswerable question answering: Retrieval coverage, abstention calibration, and hallucination-proxy analysis on SQuAD 2.0. *Journal of Technology Informatics and Engineering, 4*(2), 502–520. https://doi.org/10.51903/jtie.v4i2.536
[99] Zhong, Z. S., Li, C., & Rao, H. (2026). Trajectory reliability prediction for generalist AI agents: Tool-use failure analysis and success forecasting on ZClawBench. *Journal of Technology Informatics and Engineering, 5*(1), 341–360. https://doi.org/10.51903/jtie.v5i1.539
[100] Zhong, Z. S., & Ling, S. (2024a). Improved theoretical guarantee for rank aggregation via spectral method. *Information and Inference: A Journal of the IMA, 13*(3), Article iaae020. https://doi.org/10.1093/imaiai/iaae020
[101] Zhong, Z. S., & Ling, S. (2024b). *Uncertainty quantification of spectral estimator and MLE for orthogonal group synchronization* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2408.05944
[102] Zhong, Z. S., Pan, X., & Lei, Q. (2025). Bridging domains with approximately shared features. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics* (Vol. 258, pp. 559–567). PMLR. https://proceedings.mlr.press/v258/zhong25a.html
[103] Zhong, Z. S., Wu, Q., & Mi, G. (2025). Uncertainty-aware medical image explanation cards: LLM-generated visual explanations for AI-assisted radiology interfaces. *International Journal of Graphic Design, 3*(2), 415–436. https://doi.org/10.51903/ijgd.v3i2.3616
[104] Zhou, B., Li, C., & Liu, L. (2025). Risk-calibrated patient-facing AI safety cards: A UI/UX design framework for rubric-based medical risk communication. *International Journal of Graphic Design, 3*(2), 365–380. https://doi.org/10.51903/ijgd.v3i2.3696
[105] 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
[106] Zhou, H., & Zhang, K. (2025). News-based uncertainty and macro-market fusion for VIX direction forecasting: Evidence from 2015–2024 FRED panel. *Journal of Technology Informatics and Engineering, 4*(2), 487–501. https://doi.org/10.51903/jtie.v4i2.540
[107] Zhou, S., Chen, Y., & Lee, K. (2026). Accounting-aware evidence-constrained agents for disclosure, settlement, and secondary-market risk monitoring in tokenized RWA infrastructure. *Journal of Technology Informatics and Engineering, 5*(2), 60–74. https://doi.org/10.51903/jtie.v5i2.544