Stout in Agriculture and Animal Science (SAAS)
ISSN: 2979-2835
DOI: https://doi.org/10.61424/xbzn8d61
Journal Homepage: https://stoutjournals.org/index.php/SAAS

Correctness-Trace Replay of Commercial LLMs on CROP: Paired Performance, Shortcut Risk, and Oracle Referral Bounds

Valerie Zhang
Computer Science, Northwestern University, Evanston, IL, USA
Corresponding Author: Valerie Zhang, **E-mail**: v.zhang96@outlook.com

ARTICLE INFORMATION

Submitted: April 13, 2026
Accepted: June 29, 2026
Published: July 31, 2026
Volume: 2
Issue: 1

KEYWORDS

Crop science; large language models; CROP benchmark; paired evaluation; error complementarity; selective referral; oracle routing; benchmark shortcuts.

ABSTRACT

Reliable crop-science question answering requires more than a ranking of model accuracies: it requires evidence about paired errors, benchmark artifacts, and the attainable benefit of escalation. This study reanalyzed the 5,045-question CROP benchmark and its archived item-level correctness traces for GPT-3.5, GPT-4 Turbo, Claude-3 Opus, and Qwen-max. Question content and model outcomes were analyzed as separate evidence streams because the benchmark records and outcome workbook do not share a verified item identifier. The content audit found a strong option-length cue: gold options averaged 174.38 characters, compared with 115.38 for distractors, and an earliest-longest deterministic rule answered 4,186 questions correctly (82.97%). In the paired trace, accuracy was 32.86% for GPT-3.5, 85.69% for GPT-4 Turbo, 90.01% for Claude-3 Opus, and 86.64% for Qwen-max. Cochran's Q showed substantial heterogeneity across the four systems, and all six exact McNemar comparisons remained significant after Holm correction. Claude-3 Opus corrected 331 of GPT-4 Turbo's 722 errors and 302 of Qwen-max's 674 errors, yet shared failures limited pairwise any-correct ceilings to 92.25% and 92.63%, respectively. A post-hoc oracle reached those ceilings with small referral budgets, whereas random allocation produced only gradual gains. The findings separate measured complementarity from implementable routing: binary correctness traces support paired accuracy, error overlap, rescue rates, and oracle bounds, but not confidence calibration, reasoning-process attribution, or monetary cost. The study provides an empirically grounded reliability audit and identifies the additional logging required for risk-calibrated agricultural decision support.

1. Introduction

Large language models are increasingly considered for agricultural extension, agronomic education, and technical decision support. Their appeal is straightforward: a single interface can translate scientific information into accessible answers across crops, regions, and user expertise. Yet agricultural advice is unusually sensitive to context. Cultivar, soil, weather, pest pressure, regulation, and local practice can turn a generally plausible answer into a poor recommendation. Tzachor et al. (2023) therefore frame agricultural language models as promising access tools that still require domain oversight, while Ibrahim et al. (2024) show empirically that useful general responses can coexist with weaknesses on locally specific cultivation questions. In this setting, average accuracy is necessary but insufficient. A credible evaluation must also reveal which systems fail together, whether a stronger system can rescue another system's errors, and whether the benchmark itself rewards superficial cues.

CROP was introduced to support this kind of domain evaluation through a crop-science instruction corpus and a 5,045-question multiple-choice benchmark (Zhang et al., 2024). It belongs to a growing family of specialized agricultural evaluations. AgXQA focuses on advanced agricultural-extension question answering (Kpodo et al., 2024); SeedBench evaluates workflows in seed science (Ying et al., 2025); and AgriEval extends Chinese agricultural assessment across knowledge categories and cognitive levels (Yan et al., 2026). These benchmarks improve ecological validity relative to broad knowledge tests, but specialization does not remove measurement risk. Benchmark results can be distorted by answer-position imbalance, option-length artifacts, duplicated wording, or level definitions that depend on the same models later compared.

The natural operational response to imperfect model accuracy is selective escalation. A low-cost or incumbent model answers routine cases, while uncertain cases are sent to a stronger model or a human expert. Learned deferral methods formalize this complementarity (Madras et al., 2018; Mozannar & Sontag, 2020), and recent LLM routers seek favorable quality–cost trade-offs (Chen et al., 2024; Ong et al., 2025). However, a routing claim is only as strong as its logged evidence. A binary correctness trace can establish whether one archived system was correct when another was wrong. It cannot reconstruct confidence scores, rationales, token use, latency, or the behavior of a human agronomist. Treating an oracle rescue opportunity as a deployed router would therefore overstate the evidence.

This study asks four linked questions. First, what does the CROP question file reveal about composition, label balance, textual duplication, and answer-length cues? Second, how do the four archived commercial systems compare when evaluated as paired binary outcomes rather than independent accuracy estimates? Third, how much error complementarity is present, both as shared-error similarity and as directional rescue? Fourth, what upper limits can be placed on model-to-model referral when allocation is measured by the fraction of questions sent to the stronger system? The analysis answers these questions with deterministic content audits, Wilson confidence intervals, Cochran's Q, exact McNemar tests with Holm correction, joint outcome patterns, error-set Jaccard similarity, conditional rescue rates, and post-hoc oracle allocation envelopes.

The resulting contribution is deliberately evidence-centered. It reports complete evaluations on all 5,045 benchmark records in each available evidence stream, exposes a large option-length shortcut, and quantifies both the promise and the ceiling of model complementarity. It also clarifies why CROP's performance tiers are unsuitable as independent evidence of question difficulty in the same model comparison. Together, these results support a more rigorous interpretation of crop-science LLM benchmarking: strong accuracy remains valuable, but trustworthy deployment depends on benchmark validity, calibrated uncertainty, explicit escalation outcomes, and complete execution logs.

2. Literature Review

2.1 Domain-Specific Benchmarking and Trace Validity

Agricultural LLM evaluation has shifted from general demonstrations toward domain-specific knowledge and workflow tests. CROP's combination of a crop-science corpus and a three-tier multiple-choice benchmark made it possible to compare frontier systems on specialized agronomic content (Zhang et al., 2024). AgXQA likewise emphasizes agricultural-extension questions whose answers require more than general encyclopedic fluency (Kpodo et al., 2024). SeedBench incorporates tasks modeled on expert seed-science workflows (Ying et al., 2025), while AgriEval broadens Chinese agricultural assessment across factual domains and cognitive levels (Yan et al., 2026). This literature treats agricultural QA as a distinct measurement problem with concrete downstream consequences, not merely another general knowledge category.

The same movement toward domain-grounded evaluation is visible outside agriculture. Medical studies have progressed from task-specific image classification to lightweight foundation-model transfer and uncertainty-aware multimodal evaluation (Mi et al., 2025; Zhong et al., 2020), while cybersecurity work evaluates IoT traffic, DDoS mitigation, and host or log sequences under operationally meaningful failure modes (Li & Lu, 2025; Wang et al., 2024; Xin et al., 2024; Xin, 2026d, 2026h). These studies are not substitutes for crop-science evidence; their relevance is methodological. They show that a benchmark score is interpretable only when the task construction, data provenance, split logic, retained outputs, and deployment decision are all explicit. This principle motivates the present separation of CROP question-content auditing from paired correctness-trace analysis.

2.2 Reasoning, Verification, and Safety

One prominent approach to difficult questions is to spend more computation on reasoning. Chain-of-thought prompting elicits intermediate steps that can improve multistage inference (Wei et al., 2022), and a short step-by-step cue can produce zero-shot reasoning behavior without demonstrations (Kojima et al., 2022). Self-consistency samples multiple reasoning paths and aggregates their answers (Wang et al., 2023). Verification and iterative refinement pursue a related goal through checking or revision: backward verification reranks candidates by consistency with the problem (Weng et al., 2023), Self-Refine alternates model feedback with new responses (Madaan et al., 2023), and CRITIC grounds critique in external tools (Gou et al., 2024). The gains are not automatic. Huang et al. (2024) found that unaided self-correction can fail to repair errors and can damage an initially correct answer. These methods therefore require direct logs of prompts, generations, sampling, and verification decisions before their mechanisms can be evaluated. A single binary outcome per model identifies correctness under the archived direct-answer condition, but it does not reveal the reasoning process that produced the answer.

Safety-oriented work adds a second reason to retain the full decision trace. Continual red-teaming and behavior-level refusal research treats adversarial prompts as an evolving distribution rather than a fixed test set (Zheng et al., 2023; Zheng & Li, 2024). Evidence-based self-verification can reduce unsafe responses, but its value depends on observable verifier inputs, thresholds, and fallback actions (Zheng et al., 2024). Comparable audit logic appears in numerical guardrails, scientific-claim verification, and

attack-chain detection, where a final label is insufficient to determine whether source, arithmetic, temporal, or sequence constraints were respected (Bai et al., 2026; Z. Li et al., 2026; Su, Chen, & Qian, 2026). Reliability prediction for tool-using agents similarly benefits from trajectory features such as invalid actions and recovery patterns rather than an end-state score alone (Zhong et al., 2026). For crop advice, this implies that a self-verified answer should be evaluated as a logged sequence of claims, checks, evidence, and escalation decisions.

2.3 Calibration, Selective Risk, and Cost-Sensitive Decisions

Calibration research provides the probabilistic foundation for risk-aware selection. The Brier score is a proper squared-error criterion for probabilistic forecasts (Brier, 1950), while binned calibration estimators compare stated confidence with empirical accuracy (Naeini et al., 2015). Modern classifiers can be accurate yet overconfident, and post-hoc scaling may improve probability reliability without changing predicted classes (Guo et al., 2017). Similar problems occur in language-model question answering (Jiang et al., 2021). For LLMs, confidence can be elicited verbally (Tian et al., 2023), but its quality depends on the elicitation method and task (Xiong et al., 2024). These results explain why native probabilities or contemporaneous confidence statements are needed for ECE and Brier analysis. Correctness indicators alone have no probabilistic scale and cannot supply those metrics.

Applied studies reinforce that calibration is a decision property rather than a decorative metric. Credit-risk workflows combine probability calibration with uncertainty-driven rejection, distribution-free intervals, and explanations (Y. Chen et al., 2023; Jin et al., 2024b), while cost-sensitive fraud and bankruptcy studies show why aggregate accuracy can conceal asymmetric harm under class imbalance (Y. Chen et al., 2024; Jin et al., 2024a). Similar calibration-and-refusal designs have been evaluated for recruiter screening and fair credit decisions (Jin, 2025a; Xin, 2026c). Outside tabular risk modeling, calibrated late fusion, calibration-light brain–computer interfaces, uncertainty-aware medical vision–language classification, and probabilistic multi-sensor tracking all distinguish prediction quality from probability quality (Lu et al., 2026; Wu et al., 2025; Xin, 2025c, 2026e). The common implication for CROP is that a claimed expert-referral policy requires a probability or score observed before correctness is known; binary outcomes can establish an attainable ceiling but cannot establish calibration.

Selective prediction converts confidence into an action by allowing a system to abstain. El-Yaniv and Wiener (2010) formalized the relationship between coverage and risk among accepted predictions. Confidence thresholding can reduce retained-set risk as coverage falls (Geifman & El-Yaniv, 2017), and SelectiveNet integrates a learned selection function into training (Geifman & El-Yaniv, 2019). Question answering adds distributional instability: a selector calibrated on one domain may rank failures poorly after a shift in questions (Kamath et al., 2020). Agricultural deployment raises the same concern across crops, regions, and production systems. A useful selector must recognize difficult cases without relying on a tier label derived from the answer behavior of the evaluated models.

Risk control under shift has also been studied in operational forecasting and resource allocation. Calibrated uncertainty has been used for heterogeneous GPU-capacity forecasts, conformal demand envelopes, multi-horizon risk curves, and noisy-neighbor degradation models (He et al., 2023; Nie & Zheng, 2024; S. Chen et al., 2024; Zhao et al., 2025). Cross-cloud transfer and few-shot cold-start forecasting make the deployment shift explicit (He et al., 2024, 2025), while power-aware inventory planning and profit-aware admission control connect forecast errors to downstream actions (He et al., 2026; Zhao et al., 2026). Related work on GPU-memory prediction, market-regime uncertainty, and weather-sensitive bike-demand intervals further illustrates that uncertainty should be evaluated where regimes change, not only on an average test set (Wang et al., 2025; Xin, 2026g; H. Zhou & Zhang, 2025). The approximately shared-feature perspective provides a broader account of why transfer can succeed only when stable predictive structure survives the domain change (Zhong, Pan, & Lei, 2025). Crop-science selectors face analogous shifts across geography, season, cultivar, and management practice.

2.4 Evidence-Grounded Retrieval and Provenance

Retrieval-augmented generation addresses a different source of uncertainty: whether the model has sufficient, attributable evidence for an answer. Evidence-chain RAG combines hallucination detection, source attribution, and provenance explanation (C. Li et al., 2024), and deterministic variants emphasize that each generated claim should remain traceable to supporting material (Jin, 2025b). Evidence-calibrated evaluation on unanswerable questions shows why retrieval coverage and answerability calibration must be assessed together (Zhong, Chen, Zhong, & Sun, 2025). Multi-hop retrieval adds a budget dimension because compositional questions can require several documents before an answer is adequately supported (Su et al., 2025), while retrieval-quality prediction provides a way to test whether weak evidence can be recognized before generation (R. Zhang et al., 2025). These ideas are directly relevant to agricultural advice, where local extension bulletins, labels, and regulatory sources may disagree or omit the conditions needed for a safe recommendation.

Operational RAG studies demonstrate how the same evidence requirements can be implemented in trace-rich environments. Cloud and incident systems have paired private runbooks with hallucination-resistant answers, bilingual root-cause summaries, and multi-source attribution (Liu et al., 2023, 2024; B. Zhang et al., 2024). Retrieval-summary integration has also been used to keep code intelligence findable and explainable (Y. Li, 2024). Log-analysis work retrieves normal prototypes, creates grounded incident tickets, and constructs deterministic failure narratives rather than returning an unsupported anomaly label (Sun et al., 2026; Xin, 2025a, 2026f). Financial search, accounting review, and cross-border governance extend provenance requirements to query rewriting, disclosure evidence, settlement risk, and legal authorities (Y. Chen et al., 2025; J. Li & Zhou, 2026; Meng et al., 2026; S. Zhou et al., 2026). Review-grounded recommendation similarly evaluates whether an explanation is faithful to retrieved user evidence (Chang et al., 2026). Across these domains, provenance is not a claim that retrieval occurred; it is a retained mapping among query, source, extracted support, answer, and abstention decision.

2.5 Deferral, Routing, and Counterfactual Evaluation

Deferral extends abstention by specifying what happens next. Madras et al. (2018) treat prediction and delegation as a joint decision, Mozannar and Sontag (2020) emphasize complementarity between a model and an expert, and Okati et al. (2021) formulate triage as case-level allocation. LLM cascades apply the same principle across models. FrugalGPT showed that selective use of stronger systems can improve a quality–cost frontier (Chen et al., 2024), and RouteLLM learned a routing policy from preference data (Ong et al., 2025). A valid routing evaluation consequently needs three elements: a pre-decision signal available at inference time, observed outcomes for each destination, and a cost definition. When only paired correctness is retained, one can measure opportunity—how often a fallback was correct on base-model failures—and compute a clairvoyant ceiling. One cannot claim that a feasible selector identified those cases in advance.

Applied cascades make the resource definition concrete. Tiny-model intrusion detection and hybrid cloud–edge LLM deployment use local capacity for routine cases while reserving larger systems for harder requests (Lu & Zhou, 2024; Xin, 2025b). Cost-aware AIOps routing similarly separates parsing, detection, diagnosis, and summarization, illustrating that a single route may not be optimal for every subtask (C. Li et al., 2026). Capacity-side work forecasts demand and resource requirements before placement or admission decisions (He et al., 2023; Wang et al., 2025), and risk-aware spot-GPU control shows that routing value depends on the losses attached to rejection and overload (Zhao et al., 2026). These studies support a two-axis interpretation of the CROP envelope: referral percentage is an allocation budget, while the unobserved execution costs and harms remain separate quantities that future data collection must measure.

Offline policy evaluation offers a disciplined way to compare routing rules without confusing a retrospective optimum with a deployable policy. Counterfactual evaluation on logged recommendation data estimates policy value under explicit overlap and propensity assumptions (Mu et al., 2025), and conservative off-policy selection penalizes uncertain policy improvements (Ye et al., 2026). Those assumptions cannot be recovered from CROP's binary correctness workbook because no logging policy, propensity, or pre-decision score is present. The post-hoc oracle in this study is therefore reported only as a ceiling, while the random-allocation line supplies a transparent non-informative reference.

2.6 Human Oversight, Risk Communication, and Longitudinal Use

Human review is useful only if the interface exposes the evidence and uncertainty needed to act. Medical explanation cards combine image evidence with calibrated uncertainty, while patient-facing safety cards distinguish safe communication from mere model confidence (C. Li et al., 2025; Ye et al., 2025; Zhong, Wu, & Mi, 2025; B. Zhou et al., 2025). Scientific-search and accounting-disclosure cards organize sources, ranking cues, and visual evidence hierarchy (Jin, 2025c; K. Zhang et al., 2025). Financial and infrastructure dashboards translate forecast risk into chart and decision hierarchies, and incident cards turn distributed logs into an actionable SRE handoff (Z. Li et al., 2025; Liu et al., 2025; Nie et al., 2026). Privacy and data-integrity cards add prompt-injection and provenance checks to human oversight (Su, Rao, & Ma, 2026). For agricultural referral, an analogous expert card should show the question, candidate answer, uncertainty, cited agronomic evidence, detected conflicts, and the reason for escalation.

Evidence-linked interfaces have also been studied where the human reviewer must interpret context rather than a single label. Therapist-facing retrieval and counseling-support cards link recommendations to dialogue evidence (Y. Zhang & H. Zhang, 2025a, 2025b), while strategy-aware response control separates the intended supportive action from surface fluency (Y. Zhang & Zhou, 2026). Rubric-grounded essay scoring and student-retention systems similarly route uncertain cases to teachers while preserving criterion-level evidence and intervention rationales (Xin, 2026a, 2026b). These designs are transferable at the level of oversight: a crop expert needs reviewable grounds for intervention, not an opaque "difficult" flag.

Agricultural advisory systems may also span repeated interactions. Confidence-gated long-term memory can retain stable preferences while supporting update and forgetting (Sun et al., 2023), and federated preference learning addresses

personalization when raw histories should remain private (Kuo et al., 2025). Multilingual humanitarian RAG demonstrates that trust calibration and evidence access may vary across languages and user groups (Chen & Xu, 2026). Together with utility-preserving refusal (Zheng & Li, 2024), this work suggests that longitudinal crop assistants should remember farm constraints, cite current local evidence, and still decline or refer when evidence is missing or stale.

This distinction is especially important for interpreting model complementarity. Two models with similar aggregate accuracy can have different shared-error structures, creating unequal referral value. Conversely, a high conditional rescue rate can coexist with harmful substitutions on cases where the base model was correct and the fallback was wrong. Pairwise contingency tables, exact tests, error-set similarity, and any-correct ceilings therefore reveal information that accuracy alone obscures. The present study connects agricultural benchmark auditing to selective prediction and deferral theory while keeping observed trace results separate from the richer reasoning, calibration, and cost mechanisms described in prior work.

3. Methodology

The analysis used the complete CROP benchmark file and the complete commercial-model correctness workbook distributed with the project. The benchmark comprised 5,045 records with a question, four options, a gold answer label, and a level code. The workbook contained 5,045 rows with a source ID and one ans or err indicator for each of four systems. Table 1 records the analyzed artifacts, schemas, sizes, licenses, and SHA-256 checksums. The benchmark and code licenses were retained in the analysis package.

Table 1. Analyzed artifacts and provenance

Artifact	Records	Retained information	SHA-256
Benchmark questions	5,045	Question, four options, gold answer, level	6e54ce43a29d…
Binary correctness trace	5,045	ID and four ans/err fields	e99d098164c4…
Original correctness workbook	5,045	ID and four ans/err fields	248b4abb5a7b…
Released source archive	—	Code, prompts, results, licenses	bc557ab74b88…

Note. SHA-256 values are abbreviated here; complete values appear in the corresponding CSV file.

The two files were not joined at question level. The question records contain no item identifier, while the workbook contains no question text, answer string, or gold label. In addition, the workbook order includes source ID 109 between IDs 1366 and 1368 and omits ID 1367. Equal row counts therefore do not establish a valid positional crosswalk. Figure 1 shows the resulting design: content statistics were derived only from the benchmark JSON, and paired model statistics were derived only from rows of the outcome trace. Table 2 summarizes benchmark schema validation and duplicate checks.

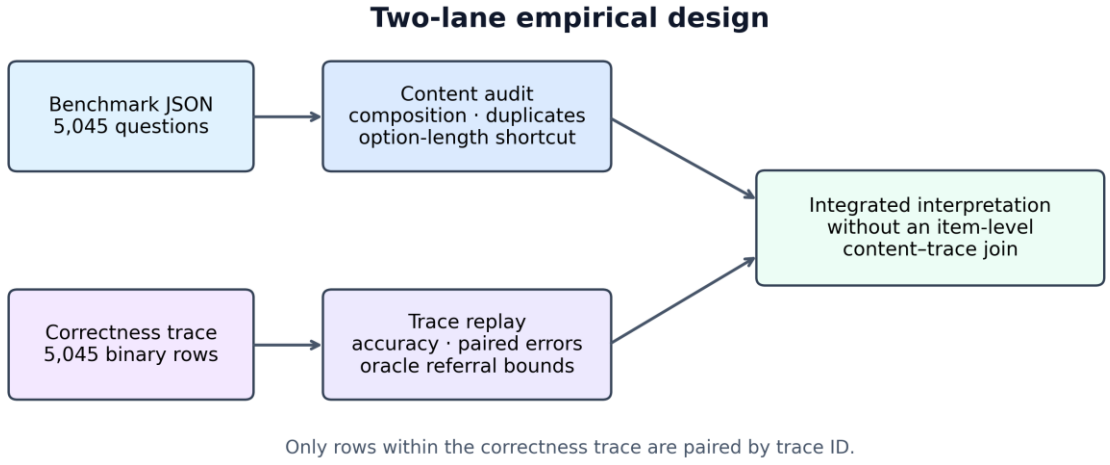

Figure 1. Two-lane empirical design.

Note. Question content and paired outcome traces are analyzed separately because no verified item-level crosswalk is present.

Table 2. Benchmark schema and duplicate screening

Metric	Value	Definition
Records	5,045	All rows passed schema validation
Fields per record	7	Question, four options, Answer, Level
Empty fields	0	All fields are non-empty strings
Unique normalized questions	5,035	NFKC, trim, case-fold, whitespace collapse
Duplicate normalized-question groups	8	Groups with at least two rows
Rows beyond first in duplicate groups	10	Ten rows across eight groups
Unique question-plus-options records	5,045	All content records are distinct
Rows with four distinct option strings	5,045	Exact string comparison

Note. Question normalization used Unicode NFKC, trimming, case folding, and whitespace collapse.

The content audit validated field completeness, answer labels, and level codes before computing descriptive statistics. Questions were assigned to the English-format group when their stored text began with the consistent leading-space convention found in the file; all remaining records formed the Chinese-format group. This rule describes file formatting rather than a linguistic classifier. Normalized question strings were created with Unicode NFKC normalization, trimming, case folding, and whitespace collapse. Duplicate groups were counted from these normalized strings, but no record was removed. Table 3 reports benchmark composition by level, format group, and gold label.

Table 3. Benchmark distribution

Dimension	Category	n	%
Level	Easy	1,613	31.97
Level	Moderate	2,710	53.72
Level	Difficult	722	14.31
Question format	Chinese format	2,523	50.01
Question format	English format	2,522	49.99
Gold answer	A	963	19.09
Gold answer	B	1,437	28.48
Gold answer	C	1,323	26.22
Gold answer	D	1,322	26.20

Note. Format groups follow the stable leading-space convention in the benchmark file; they are not language-model classifications.

The shortcut audit then measured the character length of every option after trimming. For each question, the earliest option among those tied for maximum length was selected as a deterministic baseline. A random tie-break expectation was also computed by assigning probability 1/k when the gold answer was among k longest tied options. Global and level-wise majority-label baselines provided additional content-only comparisons. Table 4 defines and reports these content measures.

Table 4. Option-length and label-only baselines

Measure	Value	Unit or rule
Mean gold-option length	174.38	characters after outer whitespace removal
Mean distractor length	115.38	characters after outer whitespace removal
Gold option among longest	87.14	percent of records
Earliest-longest heuristic accuracy	82.97	ties resolved A before B before C before D
Random longest-tie expected accuracy	83.64	uniform choice among longest tied options
Global majority-label accuracy	28.48	always predict B
Level-aware majority-label accuracy	35.38	majority label selected separately within each level

Note. Lengths were counted after outer whitespace removal. The deterministic rule resolves ties in A–D order.

The archived trace was validated independently. Each model column contained only ans and err; these values were converted to one and zero without reordering rows. The four released scripts identify the systems as gpt-3.5-turbo-0125, gpt-4-turbo-2024-04-09, claude-3-opus-20240229, and qwen_max. As shown in Table 5, the scripts requested one constrained answer and did not retain a rationale, selected option, probability, repeated sample, token count, latency, or price. GPT-3.5 and GPT-4 Turbo used temperature zero; the Claude-3 Opus and Qwen-max scripts did not set temperature explicitly.

Table 5. Archived commercial-model conditions

Model	Identifier	Condition	Temperature	Retained
GPT-3.5	gpt-3.5-turbo-0125	Single-pass, format-constrained direct answer	0	Binary correctness only
GPT-4 Turbo	gpt-4-turbo-2024-04-09	Single-pass, format-constrained direct answer	0	Binary correctness only
Claude-3 Opus	claude-3-opus-20240229	Single-pass, format-constrained direct answer	No temperature argument in released script	Binary correctness only
Qwen-max	qwen_max	Single-pass, format-constrained direct answer	API default; no temperature argument in released script	Binary correctness only

Note. Each retained result is a binary correctness indicator. The released scripts do not preserve answer strings or confidence.

Accuracy was calculated as the number of ans outcomes divided by 5,045. Each proportion received a two-sided 95% Wilson interval without continuity correction. Because all systems were evaluated on the same trace rows, model differences were tested with paired procedures. Cochran's Q assessed the omnibus null of equal correctness rates across four systems. For each unordered pair, the exact two-sided McNemar test used the two discordant cell counts. Holm's step-down method adjusted the six pairwise probabilities.

Error complementarity was measured in three ways. First, the Jaccard similarity of error sets was the number of records both models missed divided by the number missed by at least one. Second, directional rescue from base model A to comparison model B was the proportion of A's errors on which B was correct. Third, the pairwise any-correct ceiling counted a record as

solvable whenever at least one member was correct. This last quantity is an oracle bound rather than an ensemble accuracy because the trace does not reveal which option each model selected or provide a decision rule for choosing between them.

The fixed-budget referral envelope was computed from Qwen-max to Claude-3 Opus, the strongest pairwise any-correct combination. For allocation budget b, the random-allocation expectation was the linear mixture of the two marginal accuracies. The post-hoc oracle with an at-most-b budget assigned Claude first to records where Claude was correct and Qwen-max was wrong, then stopped when no further improvement was possible. Allocation was expressed as a percentage of the 5,045 records, not as monetary cost. Table 6 also reconciles the level counts. The original tier rule defines Easy from joint GPT-3.5/GPT-4 correctness, Moderate from GPT-4-only correctness, and Difficult from GPT-4 failure (Zhang et al., 2024). Because this rule is endogenous to two compared models, level was excluded from model-performance stratification and from referral decisions.

Table 6. Reconciliation of benchmark tier counts

Tier	Benchmark JSON	README	Trace rule
Easy	1,613	1,613	1,614
Moderate	2,710	2,754	2,709
Difficult	722	722	722

Note. The trace-rule reconstruction follows the tier definition reported for CROP.

All computations were executed locally with fixed deterministic rules. The analysis program validates row counts and schemas, writes item-level content and trace derivatives to separate files, produces the tables as CSV, and generates the figures from those outputs. Exact integer numerators are retained so that rounding in the paper does not affect statistical tests.

4. Results and Discussion

The benchmark contained 1,613 Easy, 2,710 Moderate, and 722 Difficult records. The format split was nearly even: 2,523 Chinese-format and 2,522 English-format questions. Gold labels were uneven, with A appearing 963 times, B 1,437 times, C 1,323 times, and D 1,322 times. Normalization found 5,035 unique question strings, including eight repeated groups and ten rows beyond the first occurrence. These repetitions were too few to dominate the aggregate results, but they are relevant to any later train–test split because near-duplicate questions should remain in the same partition.

Figure 2 and Table 4 show a much larger content artifact. Gold options averaged 174.38 characters, whereas distractors averaged 115.38. The gold answer was among the longest options on 4,396 questions (87.14%). Choosing the earliest longest option was correct on 4,186 questions, yielding 82.97% accuracy; random tie breaking among equally long maxima had an expected accuracy of 83.64%. By comparison, the global majority-label rule achieved 28.48%, and a separate majority label within each level achieved 35.38%. The length rule came within 2.72 percentage points of GPT-4 Turbo's archived accuracy. This does not show that any commercial model used the cue, because question content and traces lack a verified item crosswalk. It does show that CROP accuracy can be inflated by a deterministic surface heuristic, weakening claims that a high score alone measures crop-science reasoning.

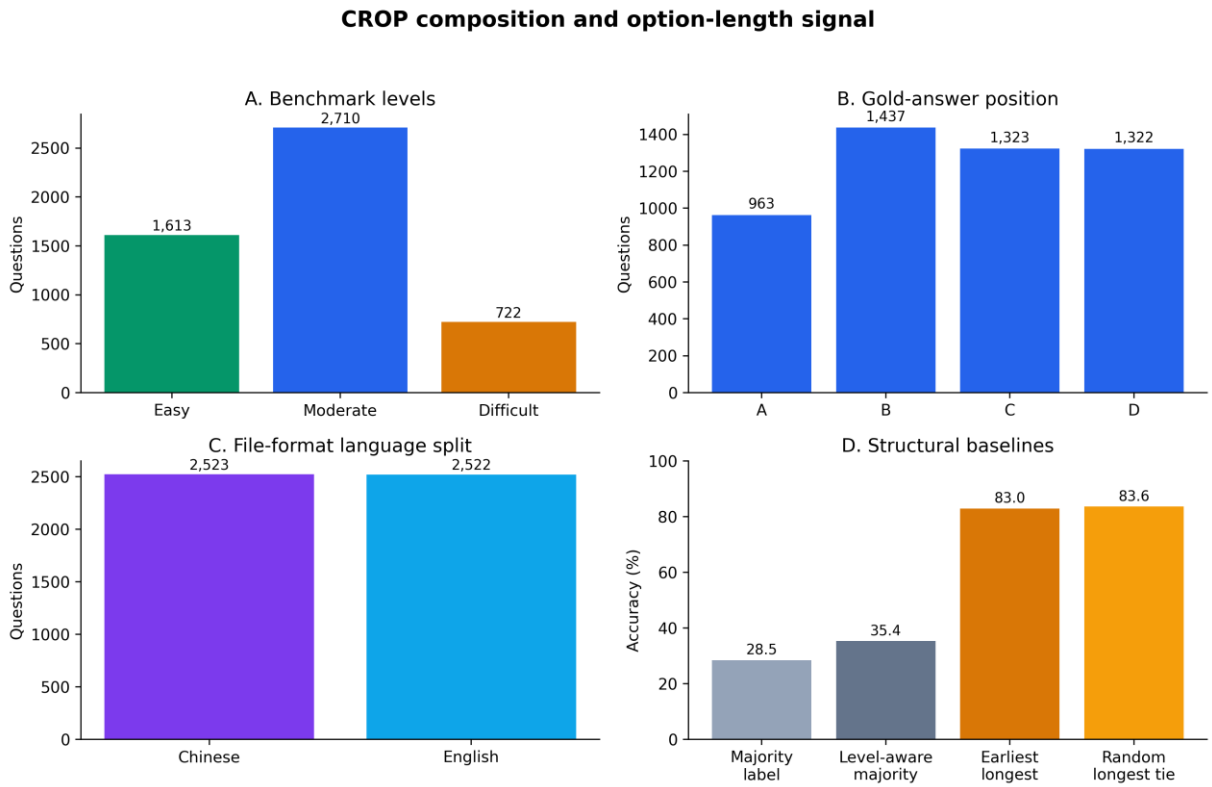

Figure 2. CROP composition and structural baselines.
Note. Panel D reports deterministic content-only baselines derived from the question file.

Table 6 identifies two further consistency issues around levels. The README count of 2,754 Moderate items would make the three reported counts sum to 5,089, whereas the benchmark contains 2,710 Moderate items. Applying the stated GPT-3.5/GPT-4 rule to the outcome trace produces 1,614 Easy, 2,709 Moderate, and 722 Difficult records, a one-record shift between Easy and Moderate relative to the JSON. More importantly, the level label is partly a restatement of GPT-3.5 and GPT-4 correctness. Reporting those models' accuracy by this level would therefore be circular: GPT-4 failure defines the Difficult group. Level-specific differences may still describe the benchmark construction, but they are not independent validation of reasoning difficulty or a legitimate feature for a deployable router.

The paired trace produced a clear performance ordering (Table 7 and Figure 3). GPT-3.5 answered 1,658 records correctly for 32.86% accuracy (95% Wilson CI [31.58%, 34.17%]). GPT-4 Turbo reached 4,323 correct, or 85.69% [84.70%, 86.63%]. Qwen-max reached 4,371 correct, or 86.64% [85.67%, 87.55%], and Claude-3 Opus led with 4,541 correct, or 90.01% [89.15%, 90.81%]. The intervals summarize uncertainty in each proportion, while the paired tests determine system differences.

Table 7. Overall archived model performance

Model	Correct	Accuracy	95% Wilson CI
GPT-3.5	1,658	32.86%	31.58%–34.17%
GPT-4 Turbo	4,323	85.69%	84.70%–86.63%
Claude-3 Opus	4,541	90.01%	89.15%–90.81%
Qwen-max	4,371	86.64%	85.67%–87.55%

Note. Intervals are two-sided 95% Wilson score intervals without continuity correction.

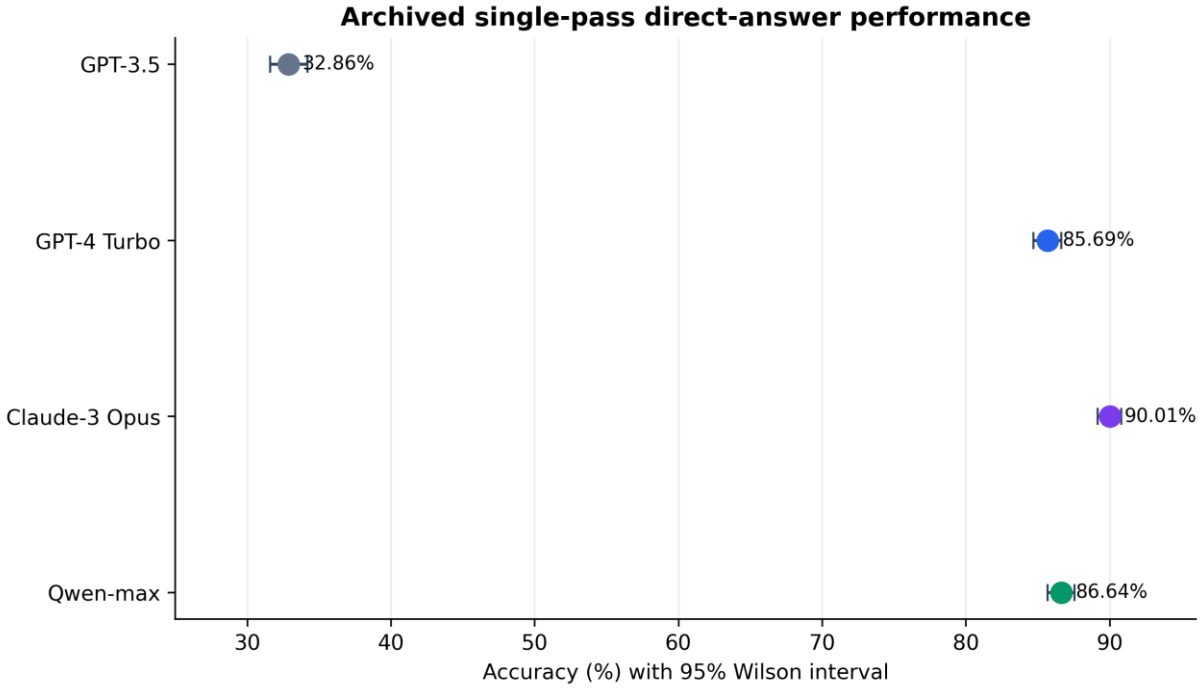

Figure 3. Archived single-pass direct-answer accuracy.
Note. Points show accuracy; whiskers show 95% Wilson intervals.

Cochran's Q was 7,013.712 with three degrees of freedom ($p < 10^{-300}$), rejecting equality across all four correctness rates. Table 8 reports the exact pairwise cells. The differences involving GPT-3.5 were especially large. Among the three stronger systems, Claude-3 Opus exceeded GPT-4 Turbo by 4.321 percentage points; the discordant counts were 331 records correct only for Claude and 113 correct only for GPT-4 Turbo, with Holm-adjusted $p = 2.066 \times 10^{-25}$. Claude exceeded Qwen-max by 3.370 points, with 302 Claude-only and 132 Qwen-only successes (Holm-adjusted $p = 4.098 \times 10^{-16}$). Qwen-max exceeded GPT-4 Turbo by 0.951 points; its 217 unique successes versus 169 for GPT-4 Turbo produced Holm-adjusted p = .01663. Pairing matters here: the tests use within-record disagreement and do not rely on overlap between marginal confidence intervals.

Table 8. Exact paired model comparisons

Model A	Model B	Δ A–B (pp)	A only	B only	Holm p
GPT-3.5	GPT-4 Turbo	-52.825	44	2,709	$< 1 \times 10^{-300}$
GPT-3.5	Claude-3 Opus	-57.146	45	2,928	$< 1 \times 10^{-300}$
GPT-3.5	Qwen-max	-53.776	36	2,749	$< 1 \times 10^{-300}$
GPT-4 Turbo	Claude-3 Opus	-4.321	113	331	2.066 × 10^-25
GPT-4 Turbo	Qwen-max	-0.951	169	217	0.01663
Claude-3 Opus	Qwen-max	3.370	302	132	4.098 × 10^-16

Note. A only and B only are the discordant correctness cells. Holm adjustment covers all six unordered pairs.

Joint correctness patterns reveal both agreement and residual headroom. All four systems were correct on 1,567 records (31.06%), and all four were wrong on 323 (6.40%). At least one system was correct on 4,722 records (93.60%). The dominant pattern was 0111: GPT-3.5 was wrong while all three stronger systems were correct, accounting for 2,514 records. Another 265 records were solved by exactly one model, 310 by exactly two, and 2,580 by exactly three. Table 9 provides all 16 observed patterns, and Figure 4 orders them by frequency. These counts are a compact checksum on the paired analysis and show that adding systems yields diminishing oracle coverage: the best single system covered 90.01%, the best pair 92.63%, the best three-system subset 93.42%, and all four 93.60%.

Table 9. Joint four-model correctness patterns

Pattern	Correct models	n	%
0111	3	2,514	49.83
1111	4	1,567	31.06
0000	0	323	6.40
0010	1	167	3.31
0011	2	137	2.72
0110	2	110	2.18
0001	1	51	1.01
0101	2	47	0.93
0100	1	38	0.75
1101	3	26	0.52
1011	3	21	0.42
1110	3	19	0.38
1000	1	9	0.18
1001	2	8	0.16
1010	2	6	0.12
1100	2	2	0.04

Note. Pattern order is GPT-3.5, GPT-4 Turbo, Claude-3 Opus, and Qwen-max; 1 indicates correct.

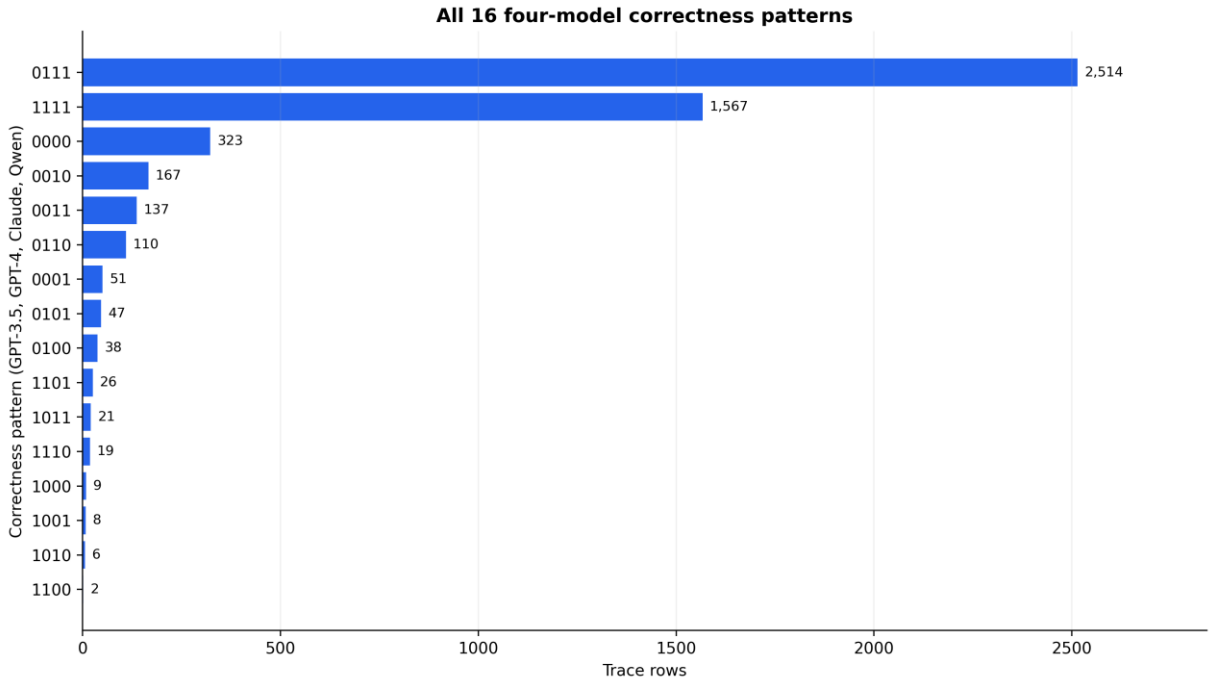

Figure 4. Complete four-model correctness-pattern distribution.
Note. Pattern positions follow GPT-3.5, GPT-4 Turbo, Claude-3 Opus, and Qwen-max.

The error-overlap analysis in Table 10 clarifies why similar accuracies do not imply interchangeable systems. GPT-4 Turbo and Qwen-max had the largest error-set Jaccard similarity, 56.68%, reflecting 505 joint errors within an 891-record error union. Claude's error Jaccard was 46.83% with GPT-4 Turbo and 46.15% with Qwen-max. Similarities involving GPT-3.5 were lower, between 13.37% and 19.76%, partly because GPT-3.5 had 3,387 errors and thus a much larger error set. Figure 5 displays the symmetric Jaccard matrix.

Table 10. Error overlap and directional rescue

Base	Alternative	Base errors	Rescued	Rescue %	Shared	Jaccard %
GPT-3.5	GPT-4 Turbo	3,387	2,709	79.98	678	19.76
GPT-3.5	Claude-3 Opus	3,387	2,928	86.45	459	13.37
GPT-3.5	Qwen-max	3,387	2,749	81.16	638	18.64
GPT-4 Turbo	GPT-3.5	722	44	6.09	678	19.76
GPT-4 Turbo	Claude-3 Opus	722	331	45.84	391	46.83
GPT-4 Turbo	Qwen-max	722	217	30.06	505	56.68
Claude-3 Opus	GPT-3.5	504	45	8.93	459	13.37
Claude-3 Opus	GPT-4 Turbo	504	113	22.42	391	46.83
Claude-3 Opus	Qwen-max	504	132	26.19	372	46.15
Qwen-max	GPT-3.5	674	36	5.34	638	18.64
Qwen-max	GPT-4 Turbo	674	169	25.07	505	56.68
Qwen-max	Claude-3 Opus	674	302	44.81	372	46.15

Note. Rescue is directional; Jaccard similarity is symmetric and uses the union of the two error sets.

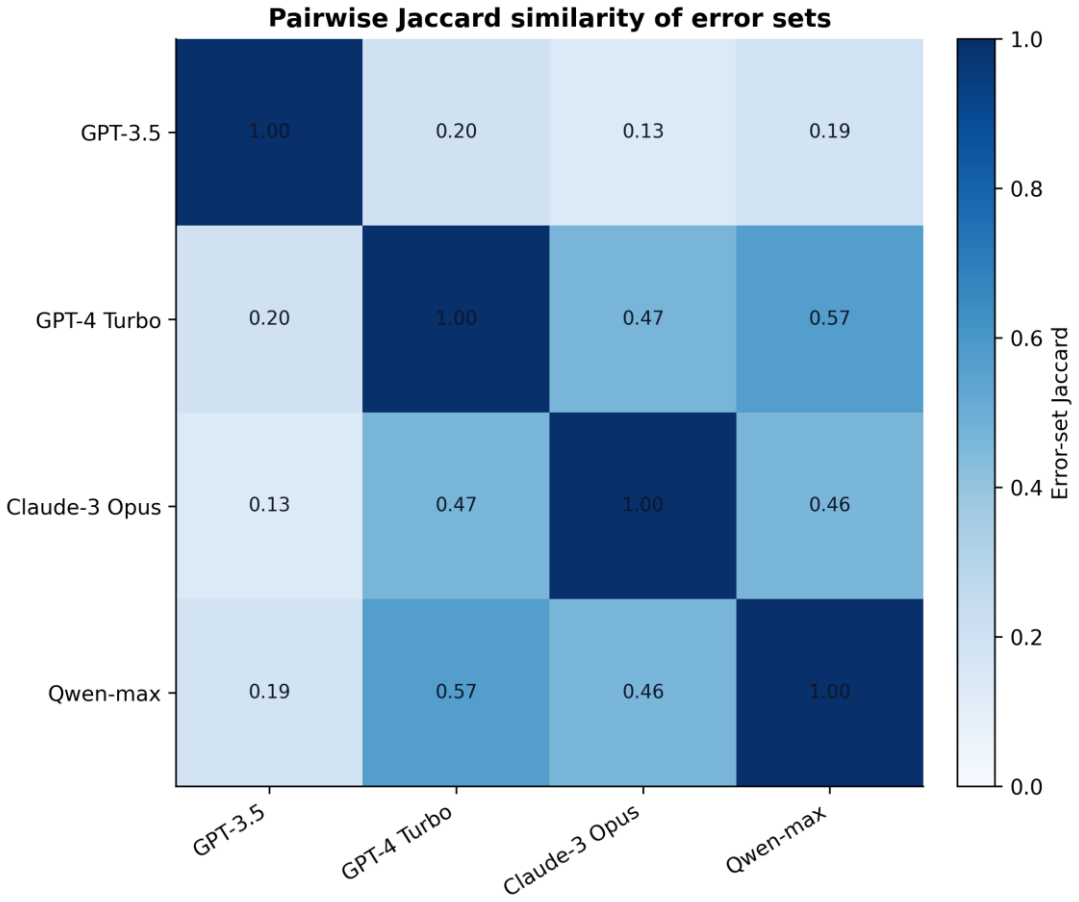

Figure 5. Pairwise Jaccard similarity of model error sets.
Note. Values are shared errors divided by the union of the two models' error sets.

Directional rescue in Figure 6 supplies the more operational view. Claude-3 Opus was correct on 331 of GPT-4 Turbo's 722 failures, a 45.84% rescue opportunity, and on 302 of Qwen-max's 674 failures, a 44.81% opportunity. Qwen-max rescued 217 of GPT-4 Turbo's failures (30.06%), while GPT-4 Turbo rescued 169 of Qwen-max's failures (25.07%). Direction matters because substitution can also cause harm: GPT-4 Turbo was correct on 113 records missed by Claude, and Qwen-max was correct on 132 such records. A routing policy must distinguish rescue cases before observing the gold label; the archived trace measures the opportunity but supplies no pre-decision signal for realizing it.

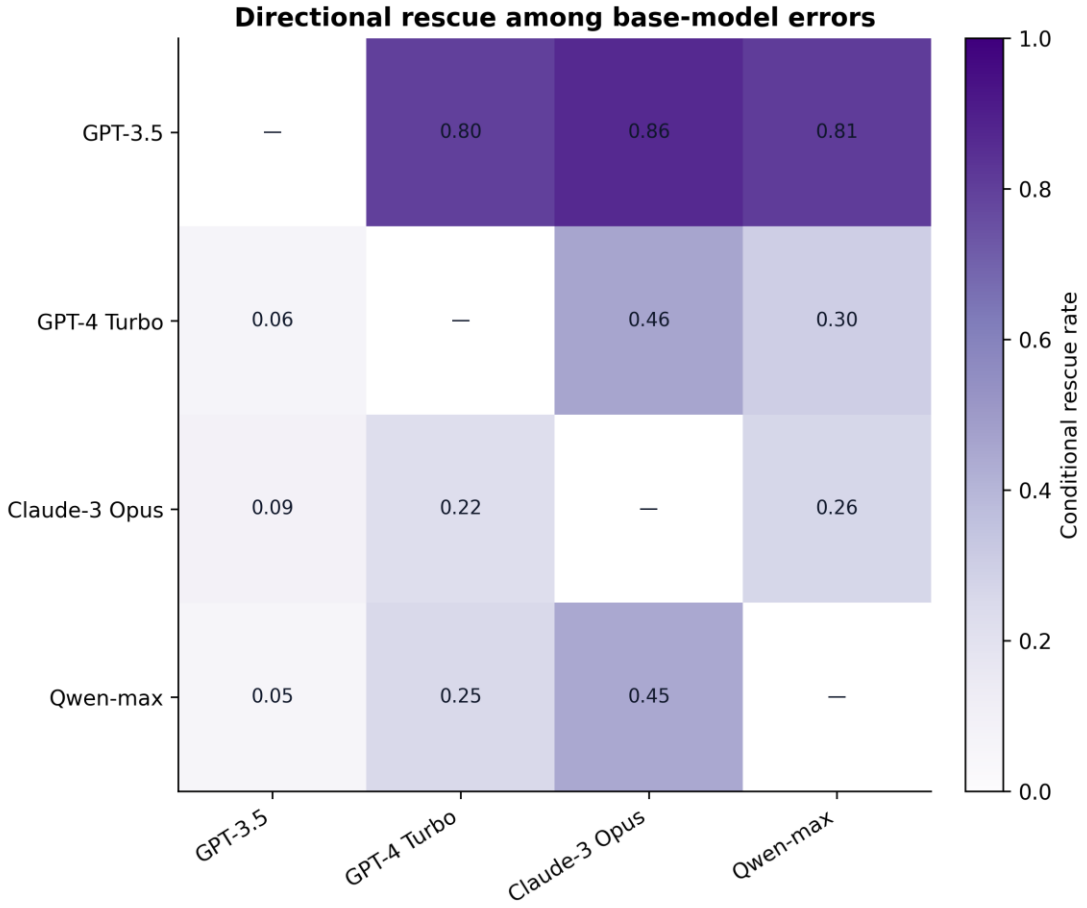

Figure 6. Directional rescue among base-model errors.
Note. Rows are base models and columns are candidate alternatives; each cell is the fraction of row-model errors corrected by the column model.

Pairwise any-correct bounds in Table 11 ranged from 80.55% for GPT-3.5 plus GPT-4 Turbo to 92.63% for Claude-3 Opus plus Qwen-max. The Claude–Qwen pair left 372 records jointly wrong and improved 2.616 percentage points over Claude alone. The GPT-4–Claude pair left 391 jointly wrong and improved 2.240 points over Claude alone. These ceilings quantify complementarity while avoiding a false ensemble claim. Because selected answers are absent, the fact that one member was correct cannot be converted into a decision rule when the systems disagree.

Table 11. Pairwise post-hoc any-correct bounds

Model pair	Both correct	A only	B only	Both wrong	Oracle correct	Oracle %
GPT-3.5 + GPT-4 Turbo	1,614	44	2,709	678	4,367	86.56
GPT-3.5 + Claude-3 Opus	1,613	45	2,928	459	4,586	90.90
GPT-3.5 + Qwen-max	1,622	36	2,749	638	4,407	87.35
GPT-4 Turbo + Claude-3 Opus	4,210	113	331	391	4,654	92.25
GPT-4 Turbo + Qwen-max	4,154	169	217	505	4,540	89.99
Claude-3 Opus + Qwen-max	4,239	302	132	372	4,673	92.63

Note. The bound counts a record as covered if either model was correct. It is not an executable ensemble.

The allocation envelope in Table 12 and Figure 7 makes the same distinction over a range of referral budgets. From Qwen-max to Claude, the oracle reaches 92.63% after allocating 302 rescue cases, 5.99% of records; random allocation at roughly 6% yields about 86.84%. The steep oracle curve is therefore an upper bound on the value of perfect error identification, not evidence that a small referral budget already achieves that score.

Table 12. Qwen-max to Claude-3 Opus allocation envelope

Budget %	Budget n	Oracle referrals	Oracle accuracy %	Random expected %
0.00	0	0	86.64	86.64
1.00	50	50	87.63	86.67
2.00	101	101	88.64	86.71
5.00	252	252	91.64	86.81
5.99	302	302	92.63	86.84
10.00	504	302	92.63	86.98
20.00	1,009	302	92.63	87.31
30.00	1,514	302	92.63	87.65
40.00	2,018	302	92.63	87.99
50.00	2,522	302	92.63	88.33
60.00	3,027	302	92.63	88.66
70.00	3,532	302	92.63	89.00
80.00	4,036	302	92.63	89.34
90.00	4,540	302	92.63	89.67
100.00	5,045	302	92.63	90.01

Note. The oracle selects only beneficial referrals and therefore represents a clairvoyant upper limit.

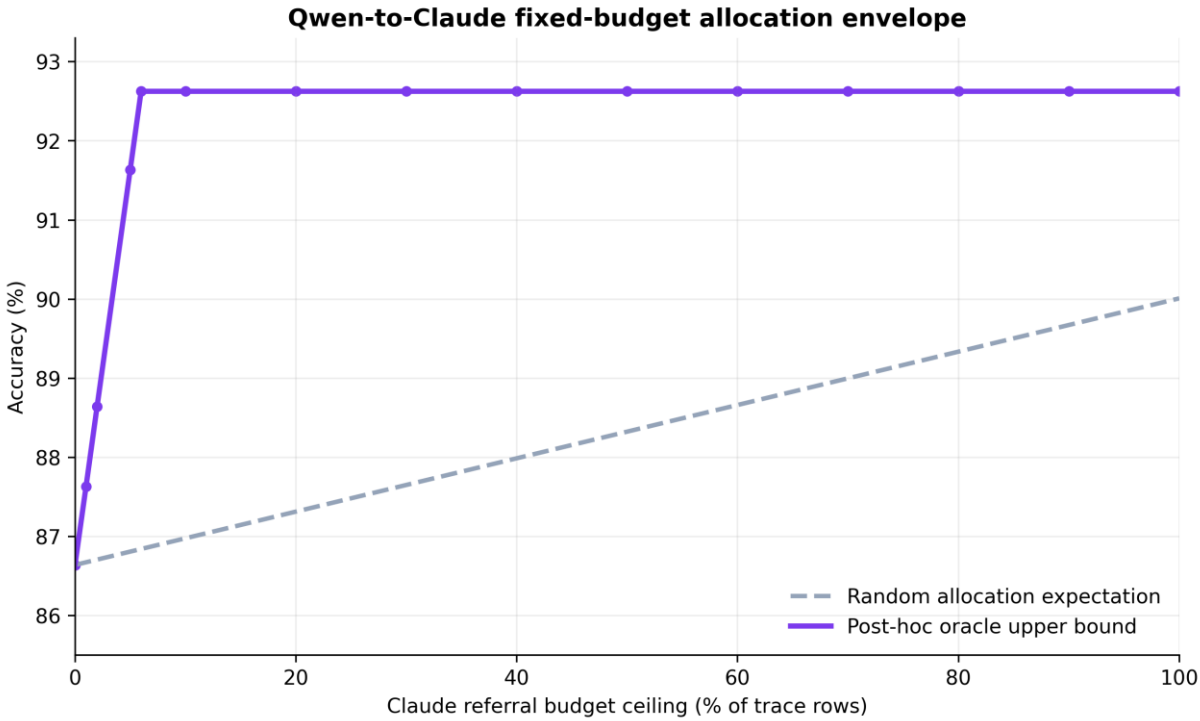

Figure 7. Qwen-max to Claude-3 Opus fixed-budget allocation envelope.
Note. The random line is an expectation; the post-hoc oracle uses correctness that is unavailable at decision time.

The contrast between the random and oracle curves defines a concrete research target. A useful verifier would need features available before the answer is judged, produce calibrated probabilities, and generalize beyond the tier construction. Its evaluation should retain native confidence or elicited confidence, compute Brier score and ECE as described by Brier (1950), Naeini et al. (2015), and Guo et al. (2017), and report risk across coverage rather than at one favorable threshold (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017). It should also log the expert or stronger-model outcome, token use, latency, and price for every route. Those records would permit genuine comparisons of direct prompting, chain-of-thought, self-consistency, verification, and routing. The present binary trace instead provides the necessary baseline: measured paired performance, observable complementarity, and the maximum gains any future selector could approach on these outcomes.

5. Conclusions

The full CROP content file and commercial-model correctness trace support a precise but bounded reliability assessment. Claude-3 Opus achieved the highest archived single-pass accuracy at 90.01%, followed by Qwen-max at 86.64%, GPT-4 Turbo at 85.69%, and GPT-3.5 at 32.86%. Exact paired tests confirmed that every model pair differed, including the comparatively small Qwen–GPT-4 gap. Complementarity was substantial but incomplete: Claude corrected roughly 45% of GPT-4 and Qwen failures, yet 372–391 errors remained shared in those pairings. Post-hoc referral could raise pairwise ceilings above 92%, but the trace contains no deployable signal that identifies rescue cases in advance.

The content audit changes how these scores should be interpreted. A simple longest-option rule reached 82.97%, demonstrating a strong surface cue, and the benchmark tiers were defined partly by GPT-3.5 and GPT-4 correctness. These properties limit claims that raw score or tier-stratified score independently measures agronomic reasoning. A next-generation CROP evaluation should balance option lengths, preserve stable item identifiers, record selected answers and rationales, collect native or elicited confidence, repeat stochastic conditions, and log execution cost. With those additions, calibration, selective risk, self-verification, and implementable routing can be tested directly. Until then, paired correctness replay offers a defensible foundation: it identifies real differences, real shared failures, and explicit upper bounds without assigning mechanisms that the evidence does not contain.

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. *Informatics, Electrical and Electronics Engineering, 6*(1), 28–43. https://doi.org/10.33474/infotron.v6i1.24923

[2] Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78*(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078%3C0001:VOFEIT%3E2.0.CO;2

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

[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/forum?id=cSimKw5p6R

[5] 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

[6] 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

[7] 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

[8] 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

[9] 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

[10] El-Yaniv, R., & Wiener, Y. (2010). On the foundations of noise-free selective classification. *Journal of Machine Learning Research, 11*(53), 1605–1641. https://jmlr.org/papers/v11/el-yaniv10a.html

[11] 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/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html

[12] Geifman, Y., & El-Yaniv, R. (2019). SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning* (PMLR Vol. 97, pp. 2151–2159). PMLR. https://proceedings.mlr.press/v97/geifman19a.html

[13] 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 *The Twelfth International Conference on Learning Representations*. https://proceedings.iclr.cc/paper_files/paper/2024/hash/fef126561bbf9d4467dbb8d27334b8fe-Abstract-Conference.html

[14] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning* (PMLR Vol. 70, pp. 1321–1330). PMLR. https://proceedings.mlr.press/v70/guo17a.html

[15] 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

[16] 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

[17] 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

[18] 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

[19] Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A. W., Song, X., & Zhou, D. (2024). Large language models cannot self-correct reasoning yet. In *The Twelfth International Conference on Learning Representations*. https://proceedings.iclr.cc/paper_files/paper/2024/hash/8b4add8b0aa8749d80a34ca5d941c355-Abstract-Conference.html

[20] Ibrahim, A., Senthilkumar, K., & Saito, K. (2024). Evaluating responses by ChatGPT to farmers' questions on irrigated lowland rice cultivation in Nigeria. *Scientific Reports, 14*, Article 3407. https://doi.org/10.1038/s41598-024-53916-1

[21] Jiang, Z., Araki, J., Ding, H., & Neubig, G. (2021). How can we know when language models know? On the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics, 9*, 962–977. https://doi.org/10.1162/tacl_a_00407

[22] Jin, J. (2025a). Calibrated resume-job matching for trustworthy LLM-assisted recruiter screening: Pairwise matching, probability calibration, and selective refusal on two public recruitment datasets. *Journal of Technology Informatics and Engineering, 4*(3), 625–648. https://doi.org/10.51903/jtie.v4i3.529

[23] Jin, J. (2025b). 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

[24] Jin, J. (2025c). 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

[25] 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

[26] 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

[27] Kamath, A., Jia, R., & Liang, P. (2020). Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5684–5696). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.503

[28] Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems, 35*, 22199–22213. https://proceedings.neurips.cc/paper_files/paper/2022/hash/8bb0d291acd4acf06ef112099c16f326-Abstract-Conference.html

[29] Kpodo, J., Kordjamshidi, P., & Nejadhashemi, A. P. (2024). AgXQA: A benchmark for advanced agricultural extension question answering. *Computers and Electronics in Agriculture, 225*, Article 109349. https://doi.org/10.1016/j.compag.2024.109349

[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] 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

[32] 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*. Advance online publication. https://doi.org/10.51903/jtie.v5i2.538

[33] 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

[34] 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

[35] 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

[36] 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

[37] 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*. Advance online publication. https://doi.org/10.51903/jtie.v5i2.541

[38] 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

[39] 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

[40] 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

[41] 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

[42] Lu, S., Chang, X., & Zou, T. (2026). Uncertainty-aware medical vision–language classification on a lightweight MedMNIST-compatible biomedical patch benchmark. *Journal of Technology Informatics and Engineering*. Advance online publication. https://doi.org/10.51903/jtie.v5i2.530

[43] 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

[44] Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., Gupta, S., Majumder, B. P., Hermann, K., Welleck, S., Yazdanbakhsh, A., & Clark, P. (2023). Self-Refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems, 36*, 46534–46594. https://proceedings.neurips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html

[45] Madras, D., Pitassi, T., & Zemel, R. (2018). Predict responsibly: Improving fairness and accuracy by learning to defer. *Advances in Neural Information Processing Systems, 31*, 6150–6160. https://proceedings.neurips.cc/paper/2018/hash/09d37c08f7b129e96277388757530c72-Abstract.html

[46] 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

[47] 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

[48] Mozannar, H., & Sontag, D. (2020). Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning* (PMLR Vol. 119, pp. 7076–7087). PMLR. https://proceedings.mlr.press/v119/mozannar20b.html

[49] 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

[50] Naeini, M. P., Cooper, G. F., & Hauskrecht, M. (2015). Obtaining well calibrated probabilities using Bayesian binning. *Proceedings of the AAAI Conference on Artificial Intelligence, 29*(1), 2901–2907. https://doi.org/10.1609/aaai.v29i1.9602

[51] 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

[52] 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

[53] Okati, N., De, A., & Gomez-Rodriguez, M. (2021). Differentiable learning under triage. *Advances in Neural Information Processing Systems, 34*, 9140–9151. https://proceedings.neurips.cc/paper/2021/hash/4c4c937b67cc8d785cea1e42ccea185c-Abstract.html

[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] 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

[56] 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

[57] 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

[58] 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

[59] 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

[60] Tian, K., Mitchell, E., Zhou, A., Sharma, A., Rafailov, R., Yao, H., Finn, C., & Manning, C. D. (2023). Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 5433–5442). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.330

[61] 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

[62] Wang, B., He, Y., Shui, Z., Xin, Q., & Lei, H. (2024). Predictive optimization of DDoS attack mitigation in distributed systems using machine learning. In *Proceedings of the 6th International Conference on Computing and Data Science* (pp. 89–94).

[63] Wang, C., Wen, Z., Zhang, R., Xu, P., & Jiang, Y. (2025). GPU memory requirement prediction for deep learning task based on bidirectional gated recurrent unit optimization transformer. In *2025 5th International Conference on Artificial Intelligence, Virtual Reality and Visualization*. IEEE. https://doi.org/10.1109/AIVRV67401.2025.11350369

[64] Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E. H., Narang, S., Chowdhery, A., & Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=1PL1NIMMrw

[65] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems, 35*, 24824–24837. https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html

[66] Weng, Y., Zhu, M., Xia, F., Li, B., He, S., Liu, S., Sun, B., Liu, K., & Zhao, J. (2023). Large language models are better reasoners with self-verification. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 2550–2575). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.167

[67] 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

[68] 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

[69] Xin, Q. (2025b). Hybrid cloud architecture for efficient and cost-effective large language model deployment. *Journal of Information Systems and Informatics, 7*(3), 2182–2195. https://doi.org/10.51519/journalisi.v7i3.1170

[70] Xin, Q. (2025c). 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). *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. *AVITEC*. Advance online publication. 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. *AVITEC*. Advance online publication. 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*. Advance online publication. https://doi.org/10.32866/001c.157499

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

[79] Xin, Q., Xu, Z., Guo, L., Zhao, F., & Wu, B. (2024). IoT traffic classification and anomaly detection method based on deep autoencoders. *Applied and Computational Engineering, 69*, 64–70. https://doi.org/10.54254/2755-2721/69/20241511

[80] Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., & Hooi, B. (2024). Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=gjeQKFxFpZ

[81] 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

[82] Ye, T., Chang, X., & Zhong, E. (2025). Uncertainty-aware breast ultrasound explanation cards: A visual communication framework for image-based AI diagnostic support using BreastMNIST_224. *International Journal of Graphic Design, 3*(2), 365–380. https://doi.org/10.51903/ijgd.v3i2.3701

[83] Ye, T., Mu, J., & Hunter, J. (2026). Off-policy evaluation and conservative policy selection for slot-level dynamic bidding and ranking on the Open Bandit Dataset (small). *Journal of Technology Informatics and Engineering, 5*(1), 178–199. https://doi.org/10.51903/jtie.v5i1.503

[84] Ying, J., Chen, Z., Wang, Z., Jiang, W., Wang, C., Yuan, Z., Su, H., Kong, H., Yang, F., & Dong, N. (2025). SeedBench: A multi-task benchmark for evaluating large language models in seed science. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 31395–31449). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.1516

[85] 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

[86] Zhang, H., Sun, J., Chen, R., Liu, W., Yuan, Z., Zheng, X., Wang, Z., Yang, Z., Yan, H., Zhong, H., Wang, X., Ouyang, W., Yang, F., & Dong, N. (2024). Empowering and assessing the utility of large language models in crop science. *Advances in Neural Information Processing Systems, 37*, 52670–52722. https://doi.org/10.52202/079017-1669

[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. In *2025 5th International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology*. IEEE. https://doi.org/10.1109/CEI66465.2025.11398480

[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*. Advance online publication. https://doi.org/10.51903/jtie.v5i2.545

[94] 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

[95] 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

[96] 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

[97] Zhong, Z., Zheng, M., Mai, H., Zhao, J., & Liu, X. (2020). Cancer image classification based on DenseNet model. *Journal of Physics: Conference Series, 1651*(1), Article 012143. https://doi.org/10.1088/1742-6596/1651/1/012143

[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., 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

[101] 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

[102] 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

[103] 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

[104] 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*. Advance online publication. https://doi.org/10.51903/jtie.v5i2.544