ADX-BENCH: A BENCHMARK FOR LLMS IN ADVERTISING FORECASTING, DIAGNOSIS, AND DECISION MAKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) are increasingly used for domain-specific reasoning and decision support, yet their capabilities in real-world advertising remain insufficiently evaluated. Advertising decision making is inherently sequential: models must anticipate future outcomes, diagnose merchant-specific performance issues, select actions under explicit business constraints, and determine whether committed decisions should be revised as new information becomes available. Existing evaluations typically abstract these capabilities into isolated tasks and focus primarily on task-level correctness, providing limited insight into whether model behavior is grounded, operationally valid, and reliable throughout the decision process. To address these challenges, we introduce **ADX-Bench**, a decision-centered benchmark for evaluating LLMs in real-world advertising. ADX-Bench comprises three complementary benchmarks constructed around a shared data lineage. **Action-Conditioned Outcome Forecasting (AOF)** evaluates multivariate multi-step forecasting from historical advertising trajectories, campaign context, and observed actions. **Evidence-Grounded Diagnosis (EGD)** evaluates taxonomy-constrained root-cause identification, merchant-specific evidence grounding, and intervention recommendation. **Constraint-Aware Action Selection (CAS)** evaluates action selection under business objectives and operational constraints, together with post-commitment reasoning when additional evidence and critiques are revealed. Built from real-world operational advertising data, ADX-Bench preserves the sequential dependencies and evolving information states of advertising decisions while enabling independent evaluation of predictive, diagnostic, and decision-making capabilities.

1 INTRODUCTION

Large language models (LLMs) are increasingly being explored beyond natural-language generation for tasks involving structured numerical data, temporal reasoning, and domain-specific decision support. Recent studies Gruver et al. (2023); Jin et al. (2024) have shown that LLMs can be adapted or directly prompted for time-series forecasting, suggesting that their learned representations can capture useful patterns beyond conventional text. In advertising, recent work Hu et al. (2026) has further begun to evaluate LLMs on realistic analytical tasks derived from professional business scenarios. However, advertising decision support requires substantially more than answering analytical questions: a model must anticipate how campaign outcomes may evolve, identify the merchant-specific factors underlying observed performance, and determine what action should be taken under explicit business constraints. These requirements arise along a sequential *forecast-diagnose-decide* decision process, where advertising decisions evolve from anticipating future outcomes, to understanding the current performance state, and ultimately to selecting and revising actions under business constraints. For forecasting, recent prompt-based and multimodal approaches Jia et al. (2024); Cao et al. (2024) have explored how pretrained language models can incorporate numerical trajectories and auxiliary information for temporal prediction. For diagnosis, recent LLM-oriented studies Xu et al. (2025) increasingly examine whether models can identify root causes and recover the evidence required to support their conclusions, showing that reliable evidence-grounded reasoning remains challenging in complex settings. For decision making, practical advertising optimization Lucier

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

et al. (2024) must account for operational requirements such as budget and return-on-investment constraints rather than simply maximizing an unconstrained objective. Although these directions have each made substantial progress, they are typically studied under task-specific formulations that abstract away the sequential dependencies and evolving decision states inherent in real-world advertising, leaving this decision process insufficiently evaluated.

Recent benchmarks have made substantial progress in evaluating capabilities relevant to advertising decision support. For forecasting, GIFT-Eval Aksu et al. (2024) provides standardized evaluation across heterogeneous time-series datasets and forecasting settings. For diagnosis and evidence reasoning, OpenRCA Xu et al. (2025) evaluates root-cause localization from heterogeneous operational signals, while DetectBench Gu et al. (2024) examines whether LLMs can identify and integrate supporting evidence from complex contexts. For interactive decision making, AgentBench Liu et al. (2024a) and WorkArena Drouin et al. (2024) evaluate reasoning and task completion in interactive and professional environments, while τ-bench Yao et al. (2024) further emphasizes domain-specific rule following and behavioral consistency. AppWorld Trivedi et al. (2024) and ToolSandbox Lu et al. (2025) introduce state-based and milestone-based evaluation for executable interactions, and CRMArena Huang et al. (2025) extends such evaluation to realistic professional workflows. Most recently, AD-Bench Hu et al. (2026) brings LLM evaluation into real-world advertising analytics through professional business requests and tool-use trajectories.

Despite this progress, two fundamental challenges remain. First, existing evaluations often abstract away the sequential structure of real-world advertising decisions. In practice, forecasting, diagnosis, action selection, and subsequent revision unfold under an evolving business state, where later decisions depend on information accumulated from earlier stages. Evaluating these capabilities in isolation therefore cannot fully capture the sequential dependencies and state transitions that characterize real advertising operations. Second, task-level correctness alone is insufficient for reliable decision support. A diagnosis may be correct but unsupported by verifiable merchant-specific evidence, while an apparently reasonable action may violate operational constraints or provide limited utility. Moreover, when new evidence or critiques are introduced, models must determine whether they genuinely justify revising a committed decision. A realistic benchmark should therefore evaluate not only output correctness, but also whether diagnoses are evidence-grounded, decisions are feasible and utility-aware, and post-commitment revisions are appropriately justified.

To address these challenges, we introduce **ADX-Bench**, a decision-centered benchmark for evaluating LLMs in real-world advertising. ADX-Bench characterizes advertising decision support through three complementary benchmarks that correspond to different stages of the underlying sequential decision process. **Action-Conditioned Outcome Forecasting (AOF)** evaluates multivariate multi-step forecasting from historical advertising trajectories, campaign context, and observed advertising actions. **Evidence-Grounded Diagnosis (EGD)** evaluates whether models can identify taxonomy-constrained root causes, ground them in verifiable merchant-specific evidence, and recommend appropriate interventions. **Constraint-Aware Action Selection (CAS)** evaluates whether models can select actions under explicit business objectives and operational constraints, and determine whether a committed decision should be retained or revised when additional evidence and critiques are revealed. ADX-Bench preserves a shared data lineage across the same underlying advertising events. Historical trajectories and observed actions define the forecasting context in AOF; corresponding merchant states are augmented with diagnostic evidence and annotations in EGD; and eligible downstream states are further instantiated as constrained decision scenarios in CAS. This design preserves the sequential dependencies and evolving information states that arise in real advertising operations while allowing each benchmark to remain independently measurable. As a result, ADX-Bench evaluates not only predictive, diagnostic, and decision-making capability, but also whether model behavior remains grounded, operationally valid, and appropriately adaptive as the decision context evolves.

Our contributions are summarized as follows:

- We introduce **ADX-Bench**, a decision-centered benchmark for evaluating LLMs across the sequential *forecast–diagnose–decide* process in real-world advertising, covering predictive, evidence-grounded, and constraint-aware decision capabilities.
- We construct three benchmarks, AOF, EGD, and CAS, around a shared data lineage that preserves the sequential dependencies among advertising trajectories, merchant-specific diagnostic evidence, and downstream decision states derived from the same underlying events.

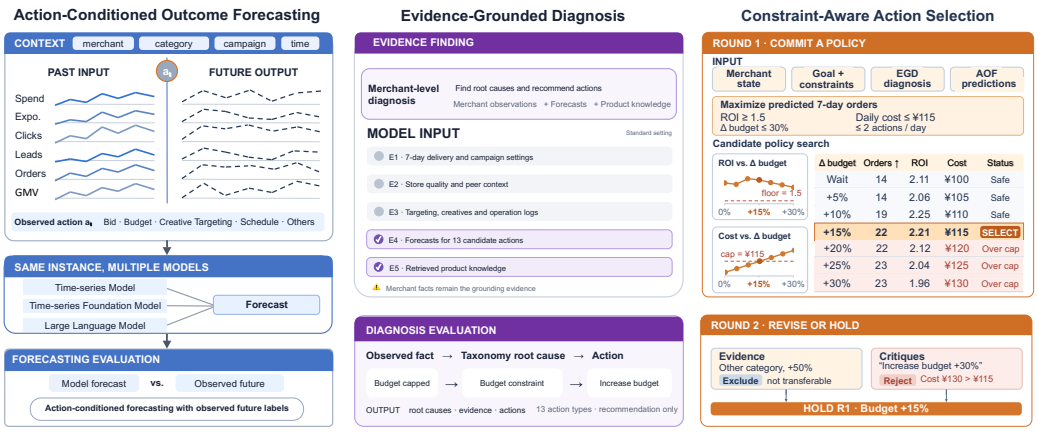


Figure 1: Architecture overview of ADX-Bench.

- We develop a reliability-oriented evaluation framework that goes beyond task-level correctness by assessing forecasting accuracy, evidence-grounded diagnosis, operationally feasible and utility-aware decisions, and justified post-commitment revision under newly revealed information.

2 Related Work

Recent LLM benchmarks evaluate capabilities beyond static question answering, including tool use, interactive reasoning, executable outcomes, and domain-specific workflows. AgentBench Liu et al. (2024a), ToolLLM Qin et al. (2024), and GAIA Mialon et al. (2024) cover general reasoning, API use, and grounded task solving. SWE-bench Jimenez et al. (2024), AppWorld Trivedi et al. (2024), and ToolSandbox Lu et al. (2025) evaluate model behavior through executable tests or state-based validation. τ-bench Yao et al. (2024) further evaluates policy compliance and final system states under repeated user–tool interactions, while WorkArena Drouin et al. (2024) and CRMArena Huang et al. (2025) extend evaluation to enterprise and professional workflows. Specialized benchmarks have also emerged for individual reasoning capabilities: GIFT-Eval Aksu et al. (2024) provides large-scale evaluation for time-series forecasting, OpenRCA Xu et al. (2025) evaluates root-cause localization from heterogeneous operational telemetry, and DetectBench Gu et al. (2024) measures whether LLMs can identify and integrate supporting evidence from complex contexts.

Advertising and recommendation benchmarks provide complementary domain-specific evaluation. MovieLens Harper & Konstan (2015) and Amazon Reviews Ni et al. (2019) support recommendation prediction and ranking, while Criteo Chapelle et al. (2014) and iPinYou Zhang et al. (2014) provide logged advertising data for response prediction and bidding evaluation. Criteo Uplift Diemert et al. (2018) supports uplift evaluation from randomized treatment data, and RecoGym Rohde et al. (2018) and RecSim Ie et al. (2019) provide simulated environments for sequential recommendation and advertising-policy evaluation. More recently, AD-Bench Hu et al. (2026) evaluates LLMs on real-world advertising analytics using professional business requests and tool-use trajectories. Despite this broad landscape, existing benchmarks largely evaluate forecasting, diagnosis, and decision making as separate capabilities.

3 ADX-Bench

Figure 1 provides an overview of ADX-Bench, which consists of three interconnected advertising benchmarks: Action-Conditioned Outcome Forecasting (AOF), Evidence-Grounded Diagnosis (EGD), and Constraint-Aware Action Selection (CAS). Rather than constructing the three benchmarks independently, ADX-Bench organizes them around a shared data lineage that connects advertising trajectories, merchant-specific diagnostic evidence, and downstream decision scenarios derived from the same underlying advertising events. This design allows each benchmark to be evaluated independently while preserving traceability across the *forecast-diagnose-decide* process.

3.1 BENCHMARK CONSTRUCTION

ADX-Bench maintains a shared data lineage across the AOF, EGD, and CAS benchmarks. Each source advertising event is assigned an anonymized lineage key derived from its campaign identifier, action timestamp, and observed action. The key is preserved across the three benchmarks, allowing forecasting instances, diagnostic records, and downstream decision scenarios to be traced back to the same underlying advertising event.

The AOF benchmark is constructed by integrating aggregated advertising-performance records, deduplicated operation logs, and merchant metadata using merchant identifiers and operation timestamps. All data are collected online from real-world advertising production systems. We restrict the benchmark to merchant actions occurring on July 15, 2026. To maintain a well-defined action-conditioned forecasting setting, we retain only merchants with exactly one recorded advertising operation on that day. Consequently, each merchant contributes exactly one merchant–action instance to AOF. Each instance corresponds to a single advertising operation and contains a 15-day temporal window, including seven days before the action day, the action day itself, and seven subsequent days. The resulting AOF benchmark contains 4,513 merchant–action instances with 217 fields, covering 12 top-level business categories and 13 observed candidate actions. These fields capture merchant characteristics, campaign configurations, advertising operations, and temporal performance signals, with merchants spanning diverse geographic regions across China. The complete feature schema and geographical distribution are provided in Appendix A.

The EGD benchmark is constructed from a subset of AOF instances with sufficient merchant-specific context and diagnostic evidence. Each instance corresponds to an anonymized advertising campaign at a fixed diagnostic date and uses frozen merchant observations inherited from AOF data. These observations include the diagnostic target and observation window, the preceding seven-day delivery trajectory, merchant quality and peer benchmarks, campaign and budget configurations, targeting and scheduling settings, creative and keyword configurations, and historical operation logs. Action-conditioned D0-D7 forecasts for 13 candidate actions and retrieved advertising product knowledge may additionally be provided as auxiliary information. Forecasts represent conditional estimates of future outcomes, whereas retrieved knowledge describes general product mechanisms and diagnostic procedures; neither is treated as merchant-specific observational evidence. EGD annotations are conducted by three senior algorithm engineers with extensive experience in advertising campaign diagnosis and optimization. Each instance is independently annotated by two experts for taxonomy-constrained root causes, recommended action types, and supporting evidence, yielding an initial agreement rate of 74%, which is computed at the instance level. Any disagreement in these annotation components is subsequently reviewed by the third expert, and the final annotation is determined by majority agreement or, when necessary, discussion among all three annotators. During annotation, experts have access to the frozen merchant-level observations described above and the shared root-cause and action taxonomies, while remaining blinded to the diagnostic outputs and evaluation scores of all evaluated models. The resulting EGD benchmark contains 100 diagnostic instances. A representative example and the root-cause distribution are provided in Appendix B.

The CAS benchmark is constructed from eligible downstream campaign states for which merchant conditions, business objectives, operational constraints, and admissible advertising actions can be explicitly specified. The resulting CAS benchmark contains 50 decision-making cases. Each instance retains the lineage of its upstream AOF forecasting record and incorporates the corresponding EGD diagnosis. CAS annotations are conducted by three senior algorithm engineers with extensive experience in advertising campaign diagnosis and optimization. Each case is independently annotated by two experts for constraint specifications, reference action plans, evidence transferability, and critique validity, yielding an initial agreement rate of 86%, which is computed at the instance level. Any disagreement in these annotation components is subsequently reviewed by the third expert, and the final annotation is determined by majority agreement or, when necessary, discussion among all three annotators. During annotation, experts have access to the merchant's preceding seven-day advertising observations and campaign configurations, business objectives and operational constraints, upstream EGD diagnoses and supporting evidence, candidate-action whitelist and AOF conditional forecasts, as well as the cross-merchant evidence pool and critiques revealed in the second round and the shared annotation guidelines. Annotators remain blinded to the action plans, revision decisions, and evaluation scores produced by all evaluated models. CAS follows a two-round delayed-revelation protocol. In round 1, the model receives the merchant state, business

objective and constraints, upstream diagnostic information, and the action whitelist. Rather than receiving all candidate-action forecasts directly, the model may request D0–D7 outcome forecasts for a limited number of candidate actions through the forecasting interface before committing to a feasible action plan. The committed plan may contain an ordered action sequence, waiting periods, stopping conditions, and rollback conditions rather than a single isolated action. In round 2, the round 1 commitment is retained, and additional evidence and critiques are revealed. The model must assess their applicability, determine whether each critique should be accepted or rejected, and decide whether the committed plan should be retained or revised. Detailed construction rules, information visibility, response formats, and representative examples are provided in Appendix C.

3.2 Evaluation Framework

ADX-Bench evaluates the three benchmarks independently while preserving their shared data lineage. Each benchmark defines its own information boundary, required model output, and evaluation criteria, allowing forecasting, diagnosis, and decision making to be assessed with metrics appropriate to their respective output spaces.

The AOF benchmark evaluates action-conditioned multivariate multi-step forecasting. Given advertising observations from D-7 to D-1, the observed advertising action, and the available campaign context, a model predicts seven advertising outcomes from D0 to D7. The observed action is used as forecasting context rather than as a randomized treatment. We use root mean squared error (**RMSE**) and mean absolute error (**MAE**) to evaluate the AOF benchmark. Each metric is first computed separately for every target variable and forecasting step over the fixed test set and then averaged across D0–D7 for each of the seven targets. To account for differences in numerical scale across targets, the resulting target-level RMSE and MAE are normalized separately by the lowest corresponding error among all evaluated models, such that the best-performing model for each target receives a normalized score of 1.000.

The EGD benchmark evaluates evidence-grounded multi-label diagnosis. Given the frozen merchant-level observations and the auxiliary information available under the corresponding evaluation setting, a model identifies root causes from a predefined taxonomy, recommends appropriate actions, and provides structured supporting evidence for each predicted root cause. Evidence references must point to verifiable merchant-specific information in the frozen input, maintaining a clear distinction among observed evidence, conditional forecasts, and general product knowledge. EGD is evaluated along three dimensions. Root-cause identification F1 (**RCF1**) measures set-level agreement between predicted and annotated root causes, while action recommendation F1 (**AF1**) measures agreement between predicted and annotated action types. Evidence support is evaluated using the evidence-grounded score (**EGS**), which considers both evidence verifiability and diagnostic support. *Validity* measures whether the referenced source, field path, and observed value can be verified against the frozen input. *Support* measures whether verified evidence directly supports a correctly identified root cause and forms a coherent diagnostic chain. Evidence entailment and diagnostic coherence are independently evaluated through repeated LLM-based judgments with majority voting; when the predicted and annotated root-cause sets have no overlap, Support is set to zero.

The CAS benchmark evaluates constraint-aware action selection and post-commitment reasoning under a two-round protocol. Round 1 (*R1 Initial*) evaluates the action plan committed before additional evidence and critiques are revealed. In Round 2 (*R2 Final*), the model assesses the newly revealed information and determines whether the committed plan should be retained or revised, after which the final plan is evaluated. CAS evaluates model performance along four dimensions on a 0-100 scale. Hard-constraint compliance (**HC**) is 100 if the model correctly determines whether a feasible action exists and its selected plan satisfies all applicable operational, numerical, and resource constraints, and 0 otherwise. Constraints requiring forecasted outcomes must be supported by an exact forecast for the selected action and parameters. Utility gain (**UG**) measures the predicted utility achieved relative to the best feasible improvement. Let u_0 denote the utility of maintaining the current configuration, u^* the best feasible utility, and u the utility of the selected feasible action. When $u^* > u_0$, we define $r = (u-u_0)/(u^*-u_0)$. UG is 100 for $r \geq 0.9$ and $100\min(1, \max(0, r))$ otherwise. When no positive feasible improvement exists, full credit is given only to a best-utility feasible decision; incorrect infeasibility judgments or unverifiable actions receive 0. Action-cost efficiency (**AE**) measures whether additional intervention cost is justified by predicted utility improvement. Low-, medium-, and high-cost actions have predefined relative costs of 1, 2, and 3,

respectively, while non-intervention has zero cost. Relative to lower-cost feasible alternatives, each additional cost unit must yield an additional predicted utility of at least 20% of the best feasible improvement; otherwise, AE is reduced proportionally to the uncompensated utility gain. Zero-cost plans or plans without a lower-cost feasible alternative receive full credit, while infeasible or unverifiable plans receive 0. Critique discrimination (**CD**) measures whether the model correctly accepts or rejects Round 2 critiques. It is computed as the macro-average recall over the accept and reject categories, with missing, contradictory, or unverified retrieval-dependent judgments counted as incorrect. HC, UG, and AE are computed per case and averaged over the 50 CAS cases for both *R1 Initial* and *R2 Final*. CD is computed across all critiques and reported only for *R2 Final*.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We evaluate AOF on three categories of forecasting methods. For conventional time-series (TS) models, we include **Informer** Zhou et al. (2021), **PatchTST** Nie et al. (2022), **NLinear** Zeng et al. (2023), **iTransformer** Liu et al. (2024b), and **TimeMixer** Wang et al. (2024). These models operate directly on structured numerical advertising trajectories. Each model is trained on the AOF training split and selected using the validation split. The observations from $D-7$ to $D-1$ are used as historical context, and the model performs multi-step forecasting for D0–D7 on the same set of target variables. For general-purpose large language models (LLMs), we evaluate **Gemini-3.1-Pro** Google DeepMind (2026), **GPT-5.6-Sol** OpenAI (2026), **DeepSeek-V4-Pro** DeepSeek-AI (2026), **Kimi-K2.6** Kimi Team (2026), **LongCat-2.0** LongCat Team (2026), and **GLM-5.2** Z.ai (2026). The historical advertising trajectory, observed advertising action, and available campaign context are converted into a standardized structured textual prompt. All LLMs receive the same input fields, prompt template, and forecasting horizon, and are instructed to generate numerical forecasts for D0–D7. Their outputs are parsed into the same structured prediction format used for evaluation. For time-series foundation models (FMs), we evaluate **Chronos-2** Ansari et al. (2025), **Sundial-base-128m** Liu et al. (2025), and **TimeMoE-50M** Shi et al. (2025). We use the publicly released pretrained checkpoints without task-specific fine-tuning and evaluate them in a zero-shot forecasting setting. Each model receives the same $D-7$ to $D-1$ historical context and predicts D0–D7 for the same target variables on the fixed test split. No AOF training examples are used to update the parameters of these foundation models. We randomly partition the 4,513 merchant-action instances into training, validation, and test sets with a 60/20/20 ratio using a fixed random seed of 42. Since each merchant contributes exactly one instance to the benchmark, the resulting partitions are mutually exclusive at the merchant level. Any data-dependent normalization statistics are estimated exclusively from the training split and subsequently applied unchanged to the validation and test splits. All baselines are evaluated under the same test split, forecasting horizon, and information boundary to ensure a consistent comparison.

We evaluate EGD and CAS on six general-purpose LLMs: **Gemini-3.1-Pro** Google DeepMind (2026), **GPT-5.6-Sol** OpenAI (2026), **DeepSeek-V4-Pro** DeepSeek-AI (2026), **Kimi-K2.6** Kimi Team (2026), **LongCat-2.0** LongCat Team (2026), and **GLM-5.2** Z.ai (2026). For EGD, all models receive the same structured diagnostic prompt, observable AOF information, taxonomy, and output schema. In forecast-augmented settings, each model additionally receives its own action-conditioned forecasts generated from AOF data as diagnostic input, preserving the model-specific information flow from AOF to EGD. For CAS, all models receive the same decision context, candidate-action whitelist, operational constraints, upstream diagnostic information, and output schema under the two-round delayed-revelation protocol. Before committing to an action plan, each model may query D0-D7 outcome forecasts for a limited number of candidate actions through the same forecasting interface.

4.2 EVALUATION ON ACTION-CONDITIONED OUTCOME FORECASTING BENCHMARK

As shown in Table 1, pretrained time-series foundation models show the clearest advantage on large-magnitude advertising signals, including Charge, Exposure, Clicks, and GMV. TimeMoE-50M achieves the best normalized performance on Charge and the lowest NRMSE on Exposure, while Sundial-base-128m performs best on Clicks and GMV and attains the lowest NMAE on Exposure. These targets correspond to relatively dense, high-volume trajectories in the original data, suggesting

Table 1: Normalized target-specific forecasting performance on the AOF benchmark for downstream large-magnitude targets.

Type	Model	Charge		Expo.		Clicks		GMV	
		NRMSE ↓	NMAE ↓	NRMSE ↓	NMAE ↓	NRMSE ↓	NMAE ↓	NRMSE ↓	NMAE ↓
TS	NLinear	1.443	1.347	1.867	1.551	1.255	1.212	1.238	1.227
	PatchTST	1.209	1.323	1.629	1.338	1.357	1.234	1.185	1.135
	Informer	1.170	1.349	1.620	1.372	1.408	1.273	1.193	1.164
	iTransformer	1.232	1.231	1.544	1.231	1.131	1.113	1.138	1.098
	TimeMixer	1.153	1.100	1.084	1.064	1.036	1.028	1.062	1.025
LLM	DeepSeek-V4-Pro	1.404	1.068	1.258	1.069	1.282	1.090	1.244	1.128
	Kimi-K2.6	2.233	1.256	1.119	1.057	1.201	1.119	1.197	1.098
	Gemini-3.1-Pro	1.425	1.159	1.093	1.042	1.166	1.092	1.190	1.109
	LongCat-2.0	1.301	1.144	1.200	1.102	1.287	1.179	1.207	1.112
	GLM-5.2	1.221	1.097	1.203	1.082	1.263	1.137	1.198	1.095
	GPT-5.6-Sol	1.246	1.143	1.199	1.074	1.272	1.141	1.172	1.069
FM	Chronos-2	1.039	1.052	1.096	1.088	1.144	1.101	1.108	1.067
	Sundial-base-128m	1.024	1.006	1.002	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	TimeMoE-50M	**1.000**	**1.000**	**1.000**	1.009	1.005	1.003	1.004	1.023

Table 2: Normalized target-specific forecasting performance on the AOF benchmark for downstream low-magnitude targets.

Type	Model	Leads		Orders		Conv.	
		NRMSE ↓	NMAE ↓	NRMSE ↓	NMAE ↓	NRMSE ↓	NMAE ↓
TS	NLinear	1.027	1.076	1.054	1.361	1.090	1.025
	PatchTST	1.028	1.050	1.028	1.036	1.096	1.006
	Informer	1.027	1.060	1.031	1.083	1.164	1.024
	iTransformer	1.022	1.068	**1.000**	1.086	1.096	**1.000**
	TimeMixer	**1.000**	**1.000**	1.082	1.057	1.104	1.018
LLM	DeepSeek-V4-Pro	1.166	1.112	1.046	**1.000**	1.323	1.086
	Kimi-K2.6	1.131	1.066	1.130	1.110	1.231	1.070
	Gemini-3.1-Pro	1.152	1.080	1.180	1.272	1.103	1.066
	LongCat-2.0	1.163	1.077	1.169	1.232	1.206	1.084
	GLM-5.2	1.150	1.068	1.161	1.174	1.229	1.073
	GPT-5.6-Sol	1.120	1.034	1.120	1.096	1.204	1.045
FM	Chronos-2	1.142	1.063	1.140	1.120	1.211	1.099
	Sundial-base-128m	1.022	1.017	1.055	1.050	**1.000**	1.039
	TimeMoE-50M	1.029	1.043	1.059	1.153	1.039	1.052

that pretrained temporal representations are particularly effective at modeling large-scale numerical dynamics and controlling large forecasting deviations. In contrast, general-purpose LLMs are generally less competitive on these targets, indicating that textualized numerical reasoning remains less precise than specialized temporal modeling for high-volume advertising signals.

As shown in Table 2, the advantage of foundation models becomes less uniform for low-magnitude downstream outcomes. For Leads, Orders, and Conversion, whose raw values and forecasting errors are substantially smaller, the best results are distributed across conventional time-series models, foundation models, and LLMs. TimeMixer achieves the best performance on Leads under both normalized metrics, while iTransformer obtains the best NRMSE for Orders and the best NMAE for Conversion. Notably, DeepSeek-V4-Pro achieves the best NMAE for Orders, whereas Sundial-base-128m achieves the best NRMSE for Conversion. The disagreement between NRMSE and NMAE is particularly informative for these low-scale targets: a model may achieve low average absolute error on typical cases while remaining more sensitive to occasional larger deviations, which are penalized more heavily by RMSE. Overall, performance on these targets is more model- and metric-dependent, suggesting that forecasting difficulty is shaped not only by model architecture but also by the scale and distributional characteristics of the target variable.

4.3 Evaluation on Evidence-Grounded Diagnosis benchmark

Table 3 reports the performance of six general-purpose LLMs on EGD under different auxiliary-input settings. The configuration with both action-conditioned forecasts and retrieved product knowledge corresponds to the full EGD setting. Under this setting, GPT-5.6-Sol achieves the strongest performance across RCF1, AF1, and EGS, while Gemini-3.1-Pro and DeepSeek-V4-Pro

Table 3: Performance on the EGD benchmark under different auxiliary-input settings. Forecast and Knowledge indicate whether action-conditioned forecasts and retrieved advertising product knowledge are provided as additional model inputs, respectively. The setting with both inputs corresponds to the full EGD configuration. Average denotes the mean performance across the six evaluated LLMs. Bold values denote the best individual-model result for each metric, while italicized values denote the best average performance across auxiliary-input settings.

Forecast	Knowledge	Model	RCF1 ↑	AF1 ↑	EGS ↑
–	–	LongCat-2.0	28.91	45.11	61.22
		Kimi-K2.6	30.91	49.95	70.15
		GLM-5.2	36.36	53.59	77.08
		Gemini-3.1-Pro	39.04	52.26	76.80
		DeepSeek-V4-Pro	38.03	49.96	80.04
		GPT-5.6-Sol	40.31	53.61	86.24
		Average	35.59	50.75	75.26
✓	–	LongCat-2.0	32.34	47.77	62.26
		Kimi-K2.6	30.63	48.20	68.45
		GLM-5.2	37.57	52.98	79.08
		Gemini-3.1-Pro	39.65	52.27	78.13
		DeepSeek-V4-Pro	36.80	50.42	75.72
		GPT-5.6-Sol	40.57	55.86	85.96
		Average	36.26	51.25	74.93
–	✓	LongCat-2.0	32.27	48.99	64.89
		Kimi-K2.6	34.22	51.80	75.94
		GLM-5.2	38.02	54.18	79.30
		Gemini-3.1-Pro	39.36	52.33	79.32
		DeepSeek-V4-Pro	39.20	52.91	79.17
		GPT-5.6-Sol	38.81	55.31	87.32
		Average	36.98	*52.59*	*77.66*
✓	✓	LongCat-2.0	30.63	46.32	63.21
		Kimi-K2.6	34.51	52.77	71.34
		GLM-5.2	38.07	52.76	78.41
		Gemini-3.1-Pro	40.10	51.99	79.22
		DeepSeek-V4-Pro	40.02	52.83	79.94
		GPT-5.6-Sol	**42.15**	**58.57**	**88.19**
		Average	*37.58*	52.54	76.72

obtain the next strongest root-cause identification results. Across models, root-cause identification remains particularly challenging relative to action recommendation and evidence grounding, indicating that recovering the underlying causes from complex merchant observations is a major source of difficulty in EGD.

The different input settings further reveal how auxiliary information affects diagnostic performance. Retrieved product knowledge provides the most consistent overall benefit: compared with the setting without auxiliary inputs, the knowledge-only setting improves the average RCF1, AF1, and EGS, and achieves the highest average AF1 and EGS. Action-conditioned forecasts alone provide smaller and less consistent gains, improving average RCF1 and AF1 while slightly reducing EGS. When both sources are available, the average RCF1 further increases to its highest level, whereas AF1 and EGS remain slightly below the knowledge-only setting. These results suggest that the two sources contribute differently to diagnosis. Product knowledge supplies general advertising mechanisms and diagnostic procedures that help models relate observed merchant conditions to plausible causes and interventions, while forecasts provide conditional information about potential future outcomes under candidate actions and are therefore more indirect diagnostic signals.

The per-model results also show that additional information is not uniformly beneficial across LLMs. Some models benefit substantially from forecasts or retrieved knowledge, whereas others exhibit limited gains or even performance degradation when more context is introduced. In particular, combining forecasts with product knowledge improves root-cause identification on average but does not consistently improve downstream action recommendation or evidence support. This suggests that effective EGD requires models to distinguish merchant-specific observations from conditional future estimates and general domain knowledge, rather than treating all available information as equivalent evidence. We provide representative case-level analyses in Appendix D to further examine these model-specific behaviors and common diagnostic failure patterns.

Table 4: Performance of LLMs on CAS benchmark.

	R1 Initial			R2 Final			
Model	**HC ↑**	**UG ↑**	**AE ↑**	**HC ↑**	**UG ↑**	**AE ↑**	**CD ↑**
Kimi-K2.6	64.0	48.1	52.1	64.0	48.8	52.0	79.8
LongCat-2.0	72.0	60.3	64.1	76.0	64.3	68.1	86.2
DeepSeek-V4-Pro	86.0	70.3	77.8	86.0	70.3	77.8	**86.4**
GLM-5.2	86.0	73.2	70.0	86.0	72.6	69.8	85.0
Gemini-3.1-Pro	86.0	**82.7**	72.5	86.0	**82.7**	74.5	86.1
GPT-5.6-Sol	**90.0**	72.7	**84.8**	**90.0**	72.7	**82.8**	84.5

4.4 Evaluation on Constraint-Aware Action Selection benchmark

Table 4 compares the R1 Initial plans with the R2 Final plans. The models exhibit clearly different strengths across the four dimensions. Gemini-3.1-Pro performs best in utility gain, GPT-5.6-Sol achieves the strongest hard-constraint compliance and action-cost efficiency, while DeepSeek-V4-Pro obtains the highest critique discrimination. No single model therefore dominates all aspects of the CAS benchmark, suggesting that constraint satisfaction, utility optimization, cost-efficient action selection, and critique discrimination capture distinct decision-making capabilities.

Comparing R1 Initial with R2 Final shows that the second-round interaction is selective rather than uniformly beneficial. Some models improve on particular dimensions, while others remain stable or introduce new errors after revision. Importantly, stable aggregate scores can mask offsetting case-level changes. For example, DeepSeek-V4-Pro obtains the same average AE of 77.75 in both rounds, despite making opposite revisions on individual cases. In case `052`, the model withdraws a 5% bid increase after recognizing that both increasing the bid and maintaining the current setting are forecast to yield 103 orders; with no predicted order gain but lower action cost, AE improves from 0 to 100. In contrast, in case `021`, the model changes from waiting to a 10% bid increase after considering positive historical evidence from peer merchants, even though the frozen forecast for the focal merchant predicts 21 orders under both options. Because the additional action cost is not compensated by a verifiable forecasted order gain, AE decreases from 100 to 0. These opposite changes cancel out at the aggregate level. Here, improvement and degradation are determined by the frozen forecasts and predefined action-cost rules rather than by subsequently observed business outcomes. This illustrates why the second round serves as a stress test of post-commitment reasoning: models must determine not only how to revise a plan, but also whether newly revealed information provides sufficient justification for revision. Additional representative failure cases are provided in Appendix E, covering errors in feasibility assessment and structured action generation.

5 Conclusions

In this paper, we introduced **ADX-Bench**, a benchmark for evaluating LLMs across forecasting, diagnosis, and decision making in real-world advertising. ADX-Bench comprises three complementary benchmarks, Action-Conditioned Outcome Forecasting (AOF), Evidence-Grounded Diagnosis (EGD), and Constraint-Aware Action Selection (CAS), that capture distinct capabilities arising along the sequential *forecast-diagnose-decide* process. Built from real-world operational advertising data, ADX-Bench preserves a shared data lineage and the evolving information states underlying advertising operations, while allowing each benchmark to be evaluated independently. Beyond task-level correctness, the benchmark further examines whether forecasts are quantitatively accurate, diagnoses are supported by verifiable merchant-specific evidence, and decisions satisfy business constraints and respond appropriately to newly revealed information. Our experiments reveal distinct challenges across the three benchmarks: specialized time-series foundation models show clear advantages on large-magnitude advertising signals, while performance on low-magnitude downstream outcomes is more model- and metric-dependent, diagnostic performance varies substantially with the availability of predictive and domain-knowledge signals, and models exhibit different strengths in constraint satisfaction, utility optimization, and post-commitment reasoning.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, manuscript organization, and code-level debugging during the preparation of this work. They were also used to help inspect intermediate experimental outputs and improve the clarity and consistency of tables and descriptions. All benchmark design choices, data construction procedures, annotations, experimental configurations, quantitative results, and scientific claims were determined or verified by the authors. The authors reviewed all AI-assisted content and take full responsibility for the final manuscript. The LLMs evaluated in ADX-Bench are part of the experimental study and are distinct from the use of generative AI tools for manuscript preparation.

ETHICS STATEMENT

ADX-Bench is constructed from real-world operational advertising data collected from production systems. Merchant, campaign, and other identifying information used in the benchmark is anonymized, and the benchmark is designed to evaluate model capabilities without exposing identifiable merchant information. The EGD and CAS benchmarks include manual annotations produced by domain experts for research evaluation. ADX-Bench is intended for studying forecasting, diagnosis, and decision-making capabilities rather than for directly controlling live advertising systems. In particular, actions produced by evaluated models are assessed in an offline setting and are not automatically executed on real advertising campaigns. We recognize that advertising decision-support systems may affect merchant outcomes if deployed in practice. Therefore, model recommendations should not be used for real-world execution without appropriate human oversight, validation, and operational safeguards.

REPRODUCIBILITY STATEMENT

To support reproducibility and future research, we will publicly release the complete ADX-Bench benchmark after appropriate privacy-preserving processing, including all benchmark data, annotations, fixed data splits, evaluation scripts and prompt templates used in this work.

REFERENCES

Taha Aksu, Gerald Woo, Juncheng Liu, Xu Liu, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. Gift-eval: A benchmark for general time series forecasting model evaluation. *arXiv preprint arXiv:2410.10393*, 2024.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, et al. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025.

Defu Cao, Furong Jia, Sercan Arik, Tomas Pfister, Yixiang Zheng, Wen Ye, and Yan Liu. Tempo: Prompt-based generative pre-trained transformer for time series forecasting. In *International Conference on Learning Representations*, volume 2024, pp. 18546–18578, 2024.

Olivier Chapelle, Eren Manavoglu, and Romer Rosales. Simple and scalable response prediction for display advertising. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(4):1–34, 2014.

DeepSeek-AI. DeepSeek-V4-Pro GA Release, 2026. Closed-source (API only).

Eustache Diemert, Artem Betlei, Christophe Renaudin, and Massih-Reza Amini. A large scale benchmark for uplift modeling. In *KDD*, 2018.

Alexandre Drouin, Maxime Gasse, Massimo Caccia, Issam H Laradji, Manuel Del Verme, Tom Marty, Léo Boisvert, Megh Thakkar, Quentin Cappart, David Vazquez, et al. Workarena: How capable are web agents at solving common knowledge work tasks? *arXiv preprint arXiv:2403.07718*, 2024.

Google DeepMind. Gemini 3.1 Pro Model Card, 2026. Closed-source (API only).

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G Wilson. Large language models are zero-shot time series forecasters. *Advances in neural information processing systems*, 36:19622–19635, 2023.

Zhouhong Gu, Lin Zhang, Xiaoxuan Zhu, Jiangjie Chen, Wenhao Huang, Yikai Zhang, Shusen Wang, Zheyu Ye, Yan Gao, Hongwei Feng, et al. Detectbench: Can large language model detect and piece together implicit evidence? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 199–222, 2024.

F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015.

Lingxiang Hu, Yiding Sun, Tianle Xia, Wenwei Li, Ming Xu, Lan Xu, Siying Wang, Wei Xu, and Jie Jiang. Ad-bench: A real-world, trajectory-aware advertising analytics benchmark for llm agents. *arXiv preprint arXiv:2602.14257*, 2026.

Kung-Hsiang Huang, Akshara Prabhakar, Sidharth Dhawan, Yixin Mao, Huan Wang, Silvio Savarese, Caiming Xiong, Philippe Laban, and Chien-Sheng Wu. Crmarena: Understanding the capacity of llm agents to perform professional crm tasks in realistic environments. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3830–3850, 2025.

Eugene Ie, Chih-wei Hsu, Martin Mladenov, Vihan Jain, Sanmit Narvekar, Jing Wang, Rui Wu, and Craig Boutilier. Recsim: A configurable simulation platform for recommender systems. *arXiv preprint arXiv:1909.04847*, 2019.

Furong Jia, Kevin Wang, Yixiang Zheng, Defu Cao, and Yan Liu. Gpt4mts: Prompt-based large language model for multimodal time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 23343–23351, 2024.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues? In *International Conference on Learning Representations*, volume 2024, pp. 54107–54157, 2024.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et al. Time-llm: Time series forecasting by reprogramming large language models. In *International conference on learning representations*, volume 2024, pp. 23857–23880, 2024.

Kimi Team. Kimi K2.6, 2026. Open-source (model weights available).

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. Agentbench: Evaluating llms as agents. In *International Conference on Learning Representations*, volume 2024, pp. 52989–53046, 2024a.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *International conference on learning representations*, volume 2024, pp. 11116–11140, 2024b.

Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Sundial: A family of highly capable time series foundation models. *arXiv preprint arXiv:2502.00816*, 2025.

LongCat Team. Introducing LongCat-2.0, 2026. Open-source (model weights available).

Jiarui Lu, Thomas Holleis, Yizhe Zhang, Bernhard Aumayer, Feng Nan, Haoping Bai, Shuang Ma, Shen Ma, Mengyu Li, Guoli Yin, et al. Toolsandbox: A stateful, conversational, interactive evaluation benchmark for llm tool use capabilities. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1160–1183, 2025.

Brendan Lucier, Sarath Pattathil, Aleksandrs Slivkins, and Mengxiao Zhang. Autobidders with budget and roi constraints: Efficiency, regret, and pacing dynamics. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 3642–3643. PMLR, 2024.

Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: a benchmark for general ai assistants. In *International Conference on Learning Representations*, volume 2024, pp. 9025–9049, 2024.

Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 188–197, 2019.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

OpenAI. GPT-5.6 System Card, 2026. Closed-source (API only).

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. Toolllm: Facilitating large language models to master 16000+ real-world apis. In *International Conference on Learning Representations*, volume 2024, pp. 9695–9717, 2024.

David Rohde, Stephen Bonner, Travis Dunlop, Flavian Vasile, and Alexandros Karatzoglou. Recogym: A reinforcement learning environment for the problem of product recommendation in online advertising. *arXiv preprint arXiv:1808.00720*, 2018.

Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. Time-moe: Billion-scale time series foundation models with mixture of experts. In *International conference on learning representations*, volume 2025, pp. 34635–34667, 2025.

Harsh Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. Appworld: A controllable world of apps and people for benchmarking interactive coding agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16022–16076, 2024.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Zhang, and Jun Zhou. Timemixer: Decomposable multiscale mixing for time series forecasting. In *International conference on learning representations*, volume 2024, pp. 38626–38652, 2024.

Junjielong Xu, Qinan Zhang, Zhiqing Zhong, Shilin He, Chaoyun Zhang, Qingwei Lin, Dan Pei, Pinjia He, Dongmei Zhang, and Qi Zhang. Openrca: Can large language models locate the root cause of software failures? In *International Conference on Learning Representations*, volume 2025, pp. 84562–84590, 2025.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024.

Z.ai. GLM-5.2, 2026. Open-source (MIT License).

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 11121–11128, 2023.

Weinan Zhang, Shuai Yuan, Jun Wang, and Xuehua Shen. Real-time bidding benchmarking with ipinyou dataset. *arXiv preprint arXiv:1407.7073*, 2014.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 11106–11115, 2021.

A AOF Benchmark Details

A.1 Merchant Distribution

The geographical distribution of merchants across China is shown in Figure 2. The distribution demonstrates the broad regional coverage of the AOF benchmark.



Figure 2: Geographical distribution of merchants in the AOF benchmark across China.

A.2 Feature Schema

The dataset contains **217 fields** covering four major categories: **merchant profile information, advertising information, operation logs, and merchant performance metrics**.

Merchant profile information describes the static characteristics of each merchant, including merchant identifiers (`shop_id`, `shop_name`, and `branch_name`), business category information (`cat0_name`, `cat1_name`, and `cat2_name`), and geographic attributes such as city, district, region, address, and geographic coordinates (`glat` and `glng`). These features provide contextual information about merchant heterogeneity across different business categories and geographic regions.

Advertising information describes the corresponding advertising campaign and historical advertising effectiveness. It includes campaign-related identifiers such as `productid`, `launchid`, and `fdate`, as well as advertising-level statistics including `total_slot_consumption`, `pay_roi`, and `merchant_roi`. These features characterize the advertising context associated with each merchant-action instance.

Operation logs capture the configuration changes that define the forecasting action. Each instance records the type of advertising operation, the corresponding configuration transition, and temporal information. Specifically, these features include operation type descriptions (`systemlogtype` and `systemlogtypedesc`), previous and updated configuration values (`valuefrom`, `value`, `valuefromdesc`, and `valuedesc`), and operation timestamps (`addtime` and `_update_timestamp`). These fields describe the action applied at the forecast origin and enable models to condition future performance prediction on the observed configuration change.

Merchant performance metrics constitute the majority of the dataset and represent the temporal advertising trajectory surrounding each advertising action. For each merchant-action instance, the dataset contains a 15-day temporal window, including seven days before the action (*D-7* to *D-1*), the action day (*D0*), and seven days after the action (*D+1* to *D+7*). Each day contains both advertising condition variables and observed performance outcomes. The advertising con-

Table 5: An anonymized EGD benchmark data example. The table shows representative content for the observable facts, auxiliary inputs, required response, and hidden evaluation reference.

EGD Item	Example
Diagnostic target and window	Diagnose the advertising campaign of an automotive-service merchant. The diagnostic date is July 15, 2026, and the observation window covers the preceding seven days.
Delivery performance and peer position	Over seven days, the campaign records RMB 1,653.58 in spend, 9,632 impressions, 600 clicks, 24 redeemed orders, and RMB 2,265 in redeemed transaction value, with an ROI of 1.37. The peer percentiles for impressions, clicks, CTR, and conversions are 0.143, 0.061, 0.245, and 0.122, respectively, indicating relatively weak traffic competitiveness.
Configuration and creative snapshot	The campaign is active with a daily budget of RMB 300 and eCPC enabled. The system-selection states for square, portrait, and landscape image slots are 2, 1, and 2, respectively, while the enabled value is 3. The corresponding numbers of available creatives are 0, 3, and 4, all below the required count of 5.
Historical operations	On July 14, the logs contain a schedule modification and multiple creative additions and deletions. The diagnostic-date snapshot nevertheless shows that system-optimized image selection remains disabled for all three image slots and that the square-image slot contains no creative.
Action-conditioned forecasts	"Optimize advertising creatives" is marked as applicable. The forecast predicts seven-day increases of 5.3% in ROI, 3.0% in CTR, and 4.5% in CVR. The forecast is used only to assess whether the candidate action may produce positive effects under the given conditions; the final root-cause diagnosis must still be supported by the creative snapshot and observed performance data.
Retrieved product knowledge	A product rule explains that system-optimized image selection can choose displayed images according to performance and should be verified against the actual states of the three image slots. The rule also states that a system-optimized image issue alone should not be treated as conclusive evidence of campaign suspension or order decline.
Required response	Each evidence reference contains `source`, `field_path`, `observed_value`, `relation`, and `is_critical`. Critical evidence must come from current merchant-level observable facts. The model outputs taxonomy-constrained root causes, relations, recommended actions, and an overall judgment.
Hidden reference and scoring	The reference root cause is "system-optimized image selection disabled," with a single-cause relation; the reference action is "optimize advertising creatives." Scoring separately evaluates the root-cause and action labels, the factual validity of cited evidence, whether the evidence directly supports the diagnosed root cause, and the coherence of the diagnostic reasoning chain.

dition variables include `bidprice`, and `pricemode`, which describe the advertising configuration under which the performance is generated. The observed outcome variables include `charge` (Charge), `expo_cnt` (Expo.), `clik_cnt` (Clicks), `leads_cnt` (Leads), `yforder_cnt` (Orders), `order_receipt_qty` (Conv.), and `buy_amt_no_shop_share` (GMV), representing different stages of the advertising funnel from exposure and engagement to conversion and transaction value. In addition to daily observations, the dataset also includes aggregated temporal statistics, such as historical period-level totals, averages, and pre/post-action comparison features, which summarize merchant advertising dynamics and temporal changes.

B EGD Benchmark Details

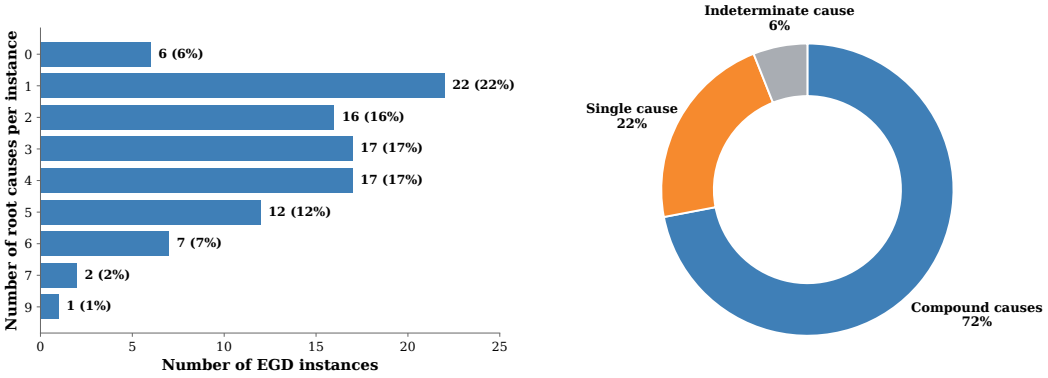
B.1 Data Example

Each EGD instance is constructed around an anonymized advertising campaign and a fixed diagnostic anchor date. All evaluation settings share exactly the same merchant-level observable facts inherited from the corresponding AOF record. These include the diagnostic target and observation window, the seven-day delivery trajectory preceding the diagnostic date, merchant quality and peer benchmarks, campaign and budget configurations, targeting and scheduling settings, creative and keyword configurations, and the corresponding operation logs. Historical operations record actions that occurred previously but are not treated as causal evidence of their effects.

The evaluation settings differ only in the availability of two auxiliary information sources: action-conditioned forecasts and retrieved advertising product knowledge. Accordingly, models are eval-

Table 6: Distribution of annotated root causes in the EGD benchmark. Coverage denotes the percentage of instances containing the corresponding root cause. Since an instance may contain multiple root causes, the percentages do not sum to 100%.

Root Cause	Count	Coverage
System-optimized image selection disabled	54	54.0%
Insufficient budget	35	35.0%
Merchant group-deal/service supply needs improvement	34	34.0%
Insufficient merchant reviews	32	32.0%
Keywords need optimization	30	30.0%
Keyword bids too low	24	24.0%
Insufficient bidding competitiveness	23	23.0%
Missing tagged images	23	23.0%
Intelligent bid adjustment disabled	19	19.0%
Low merchant rating	10	10.0%
Low merchant ROS/business rating	7	7.0%
Keyword bids not configured	6	6.0%
Restricted delivery schedule	1	1.0%



(a) Distribution of root-cause counts per instance.

(b) Distribution of indeterminate, single and compound root causes.

Figure 3: Distribution of EGD benchmark root-cause annotations.

uated with neither source, with forecasts only, with knowledge only, or with both. The forecasts provide D0–D7 conditional outcome estimates for 13 candidate actions and are used to assess the potential effects of candidate interventions under the given conditions. Retrieved knowledge describes general product mechanisms and diagnostic procedures. Neither source is treated as merchant-specific observational evidence: forecasts represent conditional future estimates rather than realized outcomes, while retrieved knowledge provides general domain information rather than facts about the focal merchant. Root-cause identification must therefore remain grounded in the merchant-level observable facts.

The model is required to output taxonomy-constrained root causes, their relations, recommended actions, and an overall diagnostic judgment. Each diagnosed root cause must be supported by structured evidence references specifying the source, field path, original observed value, and its relation to the diagnosis. Critical evidence must be grounded in the observable information from AOF. Canonical labels, reference explanations, and scoring rules remain evaluator-only and are never included in the model input. Table 5 presents a representative EGD instance.

B.2 Root-Cause Distribution

We further analyze the distribution of annotated root causes in EGD. Table 6 reports the frequency and instance-level coverage of each root-cause category. The annotations span diverse diagnostic factors, including creative configuration, merchant quality and supply, budget limitations, keyword settings, bidding competitiveness, and delivery constraints. Since an EGD instance may contain

Table 7: An anonymized CAS benchmark first-round data example. The four items constitute the complete information available to the model before it commits to an action plan.

CAS Item	Example
Merchant state	A manicure and eyelash merchant in a second-tier city: daily budget RMB 100, average daily spend RMB 97.2 over the preceding seven days (97% budget utilization), average daily orders 2.1, ROI 2.06, CPC RMB 0.81, CTR 6.7%, and CVR 1.8%. Spend and order trends are stable; intelligent bid adjustment and budget fluctuation are enabled.
Business objective and constraints	Objective: maximize order volume. Hard constraints: ROI floor 1.5, daily budget cap RMB 114, maximum per-adjustment changes of 30% for budget and 20% for bids, at most two actions per day, and no protected campaign. Soft preferences: minimize the number of actions and prefer reversible actions.
Upstream EGD diagnosis	Single root cause: "insufficient budget" with high confidence. Budget utilization reaches 97%; peer-percentile performance declines from 0.62 for impressions to 0.34 for clicks and 0.22 for orders. The recommended action is to increase the daily budget by 11%, with estimated daily gains of +2.3–6.5 clicks and +0.09–0.73 conversions.
Candidate-action whitelist and corresponding AOF forecast queries	Six `increase_budget` levels (from +5% to +30%) and `wait_and_observe` for three days. For example, the +10% option predicts daily spend of RMB 108, average daily orders of 2.4, and ROI of 2.07. The whitelist is {`increase_budget`, `wait_and_observe`, `escalate_to_human`}.

Table 8: An anonymized CAS benchmark second-round data example. The evidence pool and adversarial critiques are revealed only after the first-round commitment, while the reference response and scoring rules remain hidden.

CAS Item	Example
Evidence pool	Four items are provided: a same-category merchant with 96% budget utilization increased budget by 10%, yielding +14% orders over seven days and ROI increasing from 2.00 to 2.01; a same-category merchant increased budget by 20%, yielding +18% orders while ROI slightly decreased from 1.90 to 1.88; a same-category merchant with saturated budget waited seven days with no material metric change; and a cross-category foot-massage merchant with baseline ROI 5.5 increased budget by 50%.
Adversarial critiques	Three critiques are provided: "A 10% increase is too small; increase the budget by 30% at once"; "An impression percentile of 0.62 is acceptable, so the issue lies downstream in conversion; wait instead"; and "Enable eCPC to use the budget more efficiently."
Required response	The model outputs feasibility, an ordered action sequence with parameters and waiting periods, item-wise critique responses, evidence-counterexample assessments, modification attribution, and an overall rationale with rollback conditions. Each evidence item must be categorized as "used as support," "constitutes a counterexample," or "excluded."
Reference response and scoring	Gold response: feasible; increase budget by 10%, then wait three days. All three critiques are rejected; the cross-category evidence item is judged non-transferable and not a valid counterexample; no modification is made.

multiple root causes, the coverage rates are computed independently for each category and therefore do not sum to 100%.

Figure 3 further summarizes the structure and multiplicity of the annotations. Most instances involve compound rather than single causes, highlighting the multi-factor nature of advertising diagnosis. The number of annotated root causes also varies substantially across instances, with multi-cause cases forming the majority.

C CAS Benchmark Details

C.1 Data Example

Each CAS instance is constructed from an eligible downstream campaign state and is linked to its upstream diagnostic context through the same anonymized lineage key. The protocol is designed to evaluate post-commitment revision behavior rather than one-shot decision synthesis under full

information. Before the first-round commitment, the model receives four inputs: the merchant state, materialized from the campaign configuration at the diagnostic date and the preceding seven-day delivery trajectory; the business objective and constraints, specifying the optimization target, hard operational boundaries, and soft preferences; the upstream EGD diagnosis, providing the identified root cause, supporting evidence, and recommended intervention associated with the same lineage key; and a candidate-action whitelist defining the admissible action space, together with access to a forecasting interface through which the model may query D0-D7 conditional outcome forecasts for a limited number of candidate actions. These four inputs, including the forecasts returned through the permitted queries, constitute the complete information available before commitment. The model must then submit a decision plan containing a feasibility judgment, an ordered action sequence with execution parameters and waiting intervals, and an overall rationale including rollback conditions. The submitted plan is stored verbatim by the evaluation framework and cannot be retroactively withdrawn or rewritten. Table 7 shows the four first-round inputs for a representative CAS instance.

After the first-round commitment is fixed, the evaluation framework replays the original plan verbatim and reveals two additional inputs: a cross-merchant evidence pool and a set of adversarial critiques. The evidence pool contains a mixture of transferable, auxiliary, and deliberately non-transferable evidence items, while the critiques target specific aspects of the committed plan. The transferability labels of the evidence items and the validity labels of the critiques remain hidden from the model. The model must accept or reject each critique with justification, determine whether each evidence item constitutes a valid counterexample to the committed plan, and attribute every modification of the original plan to the evidence or critique that triggered it. Making no modification is a valid outcome when the newly revealed information does not justify revision. Each evidence item must be explicitly categorized as supporting evidence, a valid counterexample, or excluded evidence, preventing silent omission. Any revision must identify both the modified decision field and the triggering evidence or critique. Reference responses, hidden validity annotations, and deterministic scoring rules remain evaluator-only throughout evaluation. The evaluation rewards decision quality and the correctness of the model's response to newly revealed information rather than revision itself. Table 8 shows the post-commitment evidence, adversarial critiques, required response, and evaluator-only reference response and scoring for the same representative instance.

D Failure Analysis on EGD benchmark

To better understand the failure modes of current models, we conduct a paired analysis of cases that exhibit substantial performance changes under different input conditions. Overall, auxiliary information does not consistently improve diagnosis. Action-conditioned forecasts can help when they provide clear and mutually consistent signals, but may become distracting when multiple candidate actions are simultaneously predicted to be beneficial, when their implications conflict, or when the forecasts are insufficiently discriminative. The following cases illustrate these mechanisms rather than serving as substitutes for the aggregate results.

In case `042`, the annotated root cause is insufficient budget. With merchant observations only, Gemini-3.1-Pro obtains RCF1, AF1, and EGS scores of 33.33, 50.00, and 70.00, respectively; adding product knowledge improves them to 40.00, 50.00, and 100.00, indicating that retrieved knowledge helps organize a more coherent evidence chain around verifiable merchant fields. However, once action-conditioned forecasts are introduced, RCF1 and AF1 both drop to 0 and EGS decreases to 40.00, regardless of whether product knowledge is also provided. The model shifts its diagnosis toward merchant rating, review volume, conversion quality, and intelligent bid adjustment. Because several candidate interventions are predicted to yield positive outcomes, the model incorrectly infers that a potentially effective action implies the presence of its corresponding root cause. This illustrates a key distinction: action-conditioned forecasts describe potential outcomes under intervention, but do not constitute factual evidence that the corresponding root cause is currently present.

A similar pattern appears in multi-root-cause cases. Case `094` contains five annotated root causes, including low keyword bids, keywords requiring optimization, insufficient bidding competitiveness, disabled intelligent image selection, and disabled intelligent bid adjustment. With product knowledge only, Kimi-K2.6 obtains RCF1, AF1, and EGS scores of 40.00, 50.00, and 100.00; when forecasts and knowledge are provided together, these scores fall to 0, 28.57, and 40.00. The model

instead predicts restricted scheduling, weak creatives, and insufficient merchant reviews, none of which matches the annotated root causes. This case suggests that when multiple issues coexist, models must not only detect abnormalities but also determine which signals constitute primary causes. Weakly discriminative forecast signals can enlarge the hypothesis space and divert attention toward interventions with apparent predicted gains but insufficient merchant-level factual support.

Case `039` further illustrates how forecasting information can override an initially correct diagnosis. Its annotated root causes are disabled intelligent image selection, insufficient merchant reviews, and insufficient budget. With product knowledge only, LongCat-2.0 obtains 75.00 RCF1, 85.71 AF1, and 80.00 EGS, correctly identifying all three root causes and recommending largely appropriate actions. When forecasts are added, the scores decrease to 28.57, 33.33, and 66.36. The model retains only insufficient reviews while introducing restricted scheduling, low merchant rating, and missing keyword bids, despite direct observational evidence of budget saturation and inadequate creative configuration. The forecast signals therefore shift attention from observed abnormalities toward potentially beneficial interventions that do not necessarily explain the current state.

Together, these cases reveal a central limitation in current EGD models: they do not reliably distinguish the evidential roles of merchant observations, action-conditioned forecasts, and product knowledge. Merchant observations describe the current state and should provide the primary factual basis for root-cause identification. Forecasts instead characterize potential outcomes under candidate interventions and are more appropriate for evaluating whether an action may be beneficial, while product knowledge provides mechanisms and diagnostic guidance without replacing merchant-specific evidence. When these sources are not properly separated, models may reason backward from predicted action benefits to unsupported root causes or expand diagnostic hypotheses beyond what the observed evidence justifies.

This failure mode also helps explain why action recommendation can remain more stable than root-cause identification. Different or even partially incorrect diagnoses may still map to similar action types, allowing AF1 to retain partial credit even when RCF1 collapses. For example, in case `094`, RCF1 drops to 0 while AF1 remains 28.57. This indicates that a plausible action does not necessarily imply a complete or correct diagnosis. More robust models therefore need an explicit information hierarchy: first establish and verify root causes from merchant observations, then use action-conditioned forecasts to assess candidate interventions, and finally use product knowledge to explain the mechanisms linking diagnoses to actions.

E Failure Analysis on CAS benchmark

To better understand the sources of performance loss in CAS, we inspect the two-round outputs, candidate-action forecasts, and case-level scores of representative examples. The following cases illustrate two distinct failure modes involving feasibility assessment and structured action generation. All case-level scores are normalized to the range of 0-100.

In case `045`, the selected plan must simultaneously achieve at least 6.98 predicted orders over seven days, maintain an ROI of at least 4.1, and keep daily spend at most 10. GPT-5.6-Sol requests forecasts for several candidate bid increases, including 5%, 10%, and 20%. A 10% increase is predicted to yield only 6.94 orders and therefore misses the order target, while a 20% increase predicts 7 orders but reaches a maximum daily spend of 11 and an ROI of 4.0799, violating both remaining constraints. Based on the queried candidates, the model concludes that no feasible action exists and escalates the case to human intervention. However, the reference evaluation over the candidate set identifies a 15% bid increase as feasible, with 7 predicted orders, a maximum daily spend of 10, and an ROI of 4.2883. Although the second-round critique explicitly points to the 15% candidate, the model still declines to adopt it because it did not obtain the corresponding forecast itself. The failure therefore does not arise from incorrect constraint checking on the queried actions, but from generalizing from "all queried candidates are infeasible" to "no feasible candidate exists." This causes the model to lose credit on both feasibility and utility despite correctly evaluating the candidates it actually inspected.

In case `033`, all three active forecast requests issued by Kimi-K2.6 omit a recognizable tool name, leaving the model with only the baseline forecast for maintaining the current setting. In the second round, the model correctly handles all four critiques and states in its textual reasoning that

it should continue waiting and observing. However, the action field is emitted as `tool` with a leading whitespace, preventing the evaluator from recognizing the intended action. As a result, the case-level hard-constraint and utility scores both drop from 100 in Round 1 to 0 in Round 2. This failure is caused by a violation of the structured output protocol rather than by a newly selected infeasible business action. It illustrates that, in an executable decision setting, correct reasoning alone is insufficient if the resulting action cannot be represented in the required machine-readable format.

These cases are drawn from exploratory stress testing on the development set and are intended to illustrate specific failure mechanisms rather than estimate their prevalence on the held-out test set.