ANSER-BENCH: A COMPREHENSIVE BENCHMARK FOR ANALYTICAL SEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

Search and retrieval systems are increasingly used to support complex professional analysis and decision-making, where users seek not merely isolated facts but evidence-grounded conclusions to analytical questions. Meeting these needs requires retrieving heterogeneous evidence, performing multi-step aggregation and reasoning, and producing conclusions that remain traceable to their supporting evidence. Yet existing benchmarks often evaluate retrieval, question answering, and reasoning in isolation, leaving end-to-end analytical search insufficiently tested. We introduce **ANSER-Bench**,[1] a bilingual benchmark of 2,235 tasks across law, finance, and scientific research, spanning descriptive, predictive, and prescriptive analytical needs. Each task pairs a query with a domain-specific corpus, a reference conclusion, and critical-evidence annotations, enabling separate evaluation of answer quality, evidence quality, and efficiency. Our evaluation reveals substantial heterogeneity across analytical needs: across 14 task settings, no system leads in more than five, and methods that perform strongly in one setting often fail to transfer to others. This variability highlights task-dependent challenges in evidence acquisition and downstream analytical reasoning, positioning ANSER-Bench as a testbed for diagnosing such failures and advancing reliable, evidence-grounded analytical search.

1 INTRODUCTION

Search systems are increasingly expected to support not only information finding, but also complex professional analysis (Marchionini, 2006; Metzler et al., 2021). In domains such as law, finance, and scientific research, users often seek answers to questions whose conclusions may not be explicitly stated in any single document (Hou et al., 2025; Hu et al., 2026; Xie et al., 2026). For example, answering how many legal cases satisfy a set of substantive conditions may require filtering and aggregating thousands of case records; assessing whether a company's business outlook has changed may require comparing evidence across reporting periods. Such problems are referred to *analytical information needs*: information needs that require collecting, organizing, and reasoning over multiple pieces of evidence to reach an evidence-grounded conclusion (Tu et al., 2026).

Resolving analytical information needs requires substantially more effort than simply retrieving relevant documents or generating answers from short contexts (Hu et al., 2025; Luo et al., 2025). Systems must acquire sufficient evidence, integrate it through operations such as filtering, aggregation, comparison, or inference, and produce conclusions that remain verifiable against their supporting evidence. Analytical search therefore constitutes an end-to-end problem-solving process spanning *evidence retrieval*, *reasoning-intensive evidence integration*, and *conclusion verification* (Tu et al., 2026). Existing benchmarks capture important but largely separate components of this problem. Retrieval benchmarks increasingly consider complex queries but focus primarily on top-k relevance; multi-hop and retrieval-augmented generation (RAG) benchmarks couple retrieval with reasoning, yet typically operate over homogeneous textual evidence or constrained analytical patterns; research-agent benchmarks enable end-to-end information gathering and synthesis, but often target open-ended reports rather than well-defined analytical conclusions. As summarized in Table 1, among representative benchmarks, few simultaneously combine complex professional analytical needs, heterogeneous evidence acquisition, and end-to-end search-to-conclusion analysis.

[1] Benchmark is available at `https://anonymous.4open.science/r/ANSER-Bench-8947`.

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.

Table 1: Comparison with relevant benchmarks. ✓ indicates explicit coverage, ○ indicates partial coverage, and – indicates that the dimension is not a primary focus.

Benchmark	Professional Domain	Heterogeneous Evidence	Evidence Eval.	Efficiency Eval.	#Task Settings	#Query
HotpotQA (Yang et al., 2018)	–	–	✓	–	2	113K
BEIR (Thakur et al., 2021)	○	–	–	✓	9	47,033
MultiHop-RAG (Tang & Yang, 2024)	–	–	✓	–	4	2,556
MultiConIR (Lu et al., 2025)	○	–	–	–	3	23,070
STARQA (Maddela et al., 2025)	○	–	–	–	1	362
GlobalQA (Luo et al., 2025)	○	–	✓	–	4	>13K
DeepResearch Bench (Du et al., 2026)	✓	○	✓	–	1	100
FinSearchComp (Hu et al., 2026)	✓	✓	–	–	3	635
ANSER-Bench (Ours)	✓	✓	✓	✓	**14**	**2,235**

Moreover, evaluation is typically centered on final outputs, providing limited visibility into the validity of the supporting evidence and the efficiency of the overall analytical process.

To bridge these gaps, we introduce **ANSER-Bench**, a bilingual benchmark for evaluating end-to-end analytical search. It contains **2,235 tasks** across **law**, **finance**, and **scientific research**, organized around three recurring classes of analytical needs: descriptive, predictive, and prescriptive, as formally defined in Section 2.1. Each task is derived from realistic information needs and pairs a query with a domain-specific corpus from which the necessary evidence must be identified. These corpora are collected to reflect the heterogeneous information sources encountered in real-world analysis, including structured records and unstructured documents. Each task further provides a reference conclusion and annotations of critical evidence. This design enables evaluation along three complementary dimensions: whether a system reaches the correct conclusion, whether that conclusion is supported by sufficient evidence, and how efficient the overall analytical process is.

We evaluate a broad spectrum of existing approaches, including Vanilla LLM, RAG pipelines, general-purpose research agents, and domain-specialized agents, with multiple strong contemporary language-model backbones. The results reveal substantial heterogeneity across analytical needs. Across the 14 task settings, no evaluated system leads in more than five, and approaches that perform strongly in one setting frequently fail to transfer their advantage to others. At the same time, several analytical settings remain challenging for all systems: for example, the best Relaxed Accuracy@10%, measuring the proportion of answers with relative error no greater than 10%, reaches only 41.00 and 45.92 on Chinese legal and financial quantitative-aggregation tasks, respectively. Further analysis shows that higher evidence scores do not always lead to better conclusions, and higher token usage does not consistently improve answer quality. Together, these results suggest that existing methods remain highly task-dependent rather than generally capable of analytical search, with substantial performance gaps persisting in several task types and domains.

Our contributions are as follows. (1) We introduce **ANSER-Bench**, a comprehensive benchmark comprising 2,235 Chinese and English analytical tasks across law, finance, and scientific research, grounded in curated domain corpora spanning heterogeneous evidence, covering descriptive, predictive, and prescriptive analytical needs. **(2)** We develop an **evidence-grounded and efficiency-aware evaluation framework** that jointly assesses conclusion quality, critical-evidence coverage, and system efficiency, going beyond simple task accuracy to how conclusions are supported and obtained. **(3)** We **systematically evaluate representative approaches to analytical search**, spanning non-retrieval LLM, sparse and dense RAG, general-purpose and domain-specialized research agents, revealing substantial cross-task heterogeneity and the bottlenecks underlying these differences.

2 ANSER-BENCH

2.1 TASK FORMULATION AND OVERVIEW

We model the corpus available to an analytical-search system as $\mathcal{C} = \{\mathcal{D}^u, \mathcal{D}^s\}$, where $\mathcal{D}^u$ denotes unstructured evidence and $\mathcal{D}^s$ denotes structured records. A benchmark instance is represented as

$$x = (q, \mathcal{C}, E^*, y^*, m), \tag{1}$$

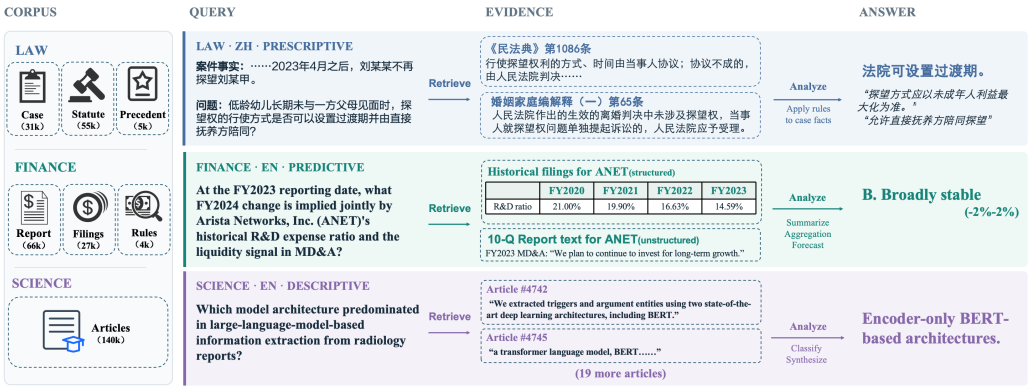


Figure 1: Illustrative ANSER-Bench corpus, query, evidence, and answer instances.

where q is an analytical query, E^* is the set of critical evidence required to derive or reproduce the reference conclusion under the specified analytical operation, y^* is the validated conclusion for the query, and m is the task metadata such as domain, language, and task type. Unlike conventional top-k relevance judgments, the granularity of E^* is task-dependent and may range from a complete target population to a small set of decisive documents, provisions, or records. Given q and access to $\mathcal{C}$, a system must identify the necessary evidence and produce an analytical conclusion.

While all tasks share the same formulation above, they differ in the analytical objective encoded by the query and the type of conclusion to be produced. Following previous studies (Tu et al., 2026), we organize analytical search tasks into three first-level categories: **Descriptive** analytical tasks focus on understanding and summarizing what has happened from observed evidence, typically through operations such as filtering, counting, ranking, aggregation, comparison and temporal analysis; **Predictive** analytical tasks move beyond summarization to infer future outcomes or latent relationships from observed evidence, including forecasting and correlation or association analysis; **Prescriptive** analytical tasks seek to determine what action, decision, obligation, or judgment should follow under specified objectives, rules, and constraints. Detailed formulations for the three task categories are provided in Appendix A.

Table 2: Task overview of ANSER-Bench. Query counts are in parentheses.

Domain–Lang.	Descriptive	Predictive	Prescriptive
Law–ZH	Case statistics (200)	Factor–outcome association (100)	Statutory rule application (100)
Law–EN	Case statistics (200)	Factor–outcome association (95)	Precedent application (57)
Finance–ZH	Filing analysis (200) Corpus statistics (100)	Financial forecasting (200)	Disclosure compliance (100)
Finance–EN	Filing analysis (200) Corpus statistics (100)	Financial forecasting (200)	Capital-rule assessment (200) Filing-fact audit (100)
Science–EN	Pattern synthesis (20)	Effect / association synthesis (63)	–

ANSER-Bench contains 2,235 analytical tasks across law, finance, and scientific research in Chinese and English, organized into 14 *task settings* defined by domain, language, and analytical need. A task setting may comprise multiple evaluation subsets; for example, Finance Descriptive includes filing analysis and corpus statistics. Table 2 summarizes the task settings and their operational subtasks, while Figure 1 illustrates representative query, evidence, and answer instances across domains. The corpus, query form, and evidence form associated with each subtask are detailed in Appendix B.

2.2 Benchmark Construction

ANSER-Bench is primarily newly constructed, with a small number of portions adapted from existing benchmarks: CLERC (Hou et al., 2025) for English legal prescriptive tasks, FinAuditing (Wang et al., 2026) and RegBench (RegBench Team, 2026) for English financial prescriptive tasks, and MetaSyn (Xie et al., 2026) for scientific tasks. We develop a human-in-the-loop four-stage construction protocol, illustrated in Figure 2; the human review procedures are described in the Quality

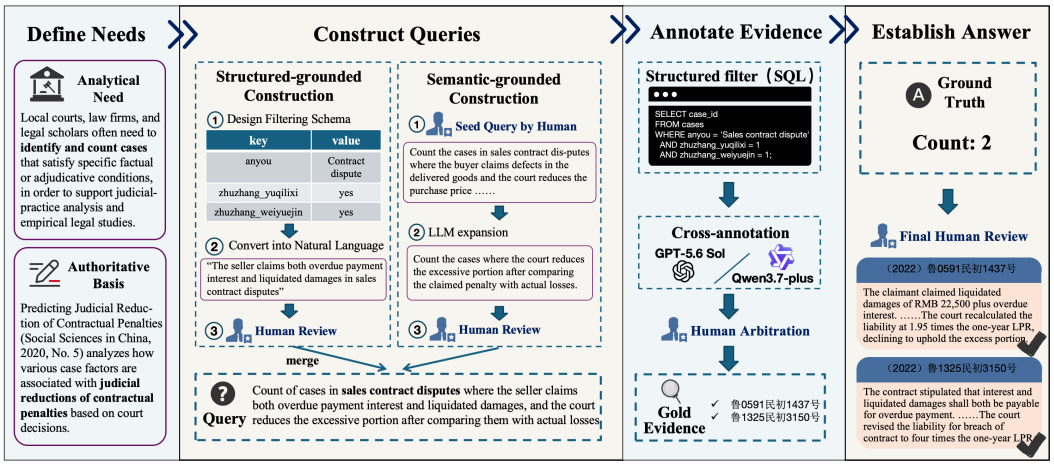


Figure 2: Illustrative construction protocol: define a professional need, dual-path query construction, screen and review supporting evidence, and establish the reference answer.

Control section below. The same protocol has also been used to adapt tasks from relevant benchmark (as described previously) to fit the needs of ANSER-Bench. Detailed domain-specific construction procedures and examples are provided in Appendices C–E.

Define needs and assemble the corpus. We first derive recurring analytical needs from academic literature, official documents, regulatory materials, and industry reports across law, finance, and scientific research, and consolidate them into task types reflecting recurring professional operations, such as legal case statistics, financial forecasting, and scientific evidence synthesis. For each need, we construct a corresponding corpus from authoritative domain sources, including judicial repositories, corporate filings, exchange and regulatory materials, and scientific articles. Unstructured documents are collected, cleaned, segmented into traceable retrieval units, and assigned stable source identifiers. To obtain task-relevant structured representations, domain experts first design schemas according to domain conventions and analytical requirements, with fields covering Boolean indicators, numerical values, categorical or list-valued attributes, and free-text strings. These fields are then populated from source documents using a combination of rule-based extraction and LLM-assisted extraction. The resulting records are subsequently checked against the original documents, corrected where necessary, normalized to the predefined schema, and linked back to their source.

Construct queries. All queries are motivated by authentic professional needs and traceable to authoritative sources that motivate their analytical targets, constraints, or query patterns. For example, Chinese legal case-statistics queries are designed with reference to the categories and analytical dimensions used in the Supreme People's Court's official judicial statistics, as published in the Gazette of the Supreme People's Court[2]. We construct queries through two complementary paths. In **structured-grounded construction**, annotators start from structured records, design executable filtering and calculation logic, obtain reproducible results through queries or scripts, and then translate the analytical conditions into natural-language queries. In **semantic-grounded construction**, domain experts derive seed queries from authoritative materials, summarize reusable templates and use LLMs to assist semantic expansion and rewriting, followed by human review. Most ANSER-Bench instances combine the two paths: the structured component specifies qualifying sets, executable constraints, and quantitative conditions, while the semantic component contributes document-level conditions and domain-specific interpretations.

Annotate evidence. For each query, we construct a gold evidence set E^* through a staged procedure. We first reuse the structured component of the query-construction logic to filter the corpus into an initial candidate pool. GPT-5.6 Sol and Qwen3.7-Plus then independently assess each candidate evidence unit against the remaining semantic conditions and evidence requirements. Disagreements between the two annotations are manually adjudicated to produce the final evidence set.

[2] https://gongbao.court.gov.cn/

Establish answers. We establish reference conclusions through five task-specific protocols, chosen according to the strongest available form of supervision. **Deterministic computation** derives answers from executable SQL, statistical, or numerical procedures for tasks such as case statistics and association analysis. **Temporal holdout** uses outcomes observed after a predefined cut-off as gold labels for forecasting tasks, while exposing only information available before that cut-off in the corpus. **Authoritative-source grounding** derives legal and regulatory judgments from official rules, decisions, precedents, or other normative records. **Systematic-review-grounded synthesis** uses validated systematic reviews or meta-analyses to construct conclusions for scientific evidence-synthesis tasks. Finally, **expert-grounded annotation** is used when conclusions cannot be uniquely determined by the preceding sources, with domain annotators independently deriving evidence-grounded answers and disagreements resolved through adjudication.

Quality control. Annotation quality is controlled through domain-specific human review. For law and finance, we recruited three annotators per domain, all second-year or more senior PhD students from law and finance schools, respectively. All legal annotators had passed China's National Unified Legal Professional Qualification Examination, while all financial annotators had passed their doctoral qualifying examination. For scientific-research tasks, we recruited three PhD students from relevant disciplines including medicine and public health, each with at least one year of research experience in meta-analysis. Before formal review, annotators follow a shared annotation manual and complete a 20-query pilot round to calibrate task definitions, evidence criteria, and answer formats. **All queries** are manually checked for validity, clarity, answerability, and consistency with the intended task, while **all conclusions** are independently re-derived by a second annotator from the query and evidence without access to the original answer. We further sample **10% of evidence annotations** for manual verification of missing, redundant, or incorrect evidence; systematic errors are traced back to tasks constructed under the same rule or filtering condition and corrected accordingly. Additional quality-control details, including the disciplinary backgrounds of annotators and the annotation focus specific to each domain, are provided in Appendices C.5, D.5 and E.3.

2.3 Evaluation

We evaluate analytical-search systems along three complementary dimensions: **answer quality**, **evidence quality**, and **efficiency**. For each task setting, we designate one primary answer metric that matches its output form and can be expressed on a bounded, higher-is-better scale for consistent reporting. Categorical and multiple-choice tasks use accuracy; numerical aggregation tasks use Relaxed Accuracy@10% (RA@10%), defined as the proportion of predictions within 10% relative error of the reference value, allowing small deviations in large-scale counts and statistics. Open-ended conclusions are evaluated by a GPT-5.6 Sol LLM judge with human validation, while scientific tasks with short, relatively constrained answers use Token-F1. Additional metrics, such as MAE and SMAPE for numerical answers, are reported as secondary diagnostics to capture complementary aspects of answer quality. Only primary answer metrics enter macro-level reporting: evaluation subsets, such as filing analysis and corpus statistics within Finance Descriptive task, are averaged equally within each task setting, after which task-setting scores are averaged equally at higher levels.

Evidence evaluation follows the required evidence granularity. For tasks whose analytical result depends on identifying the complete set of evidence units satisfying the query conditions, which we define as **population-level tasks**, we use set precision, recall, and F1. These include legal descriptive and predictive tasks and the financial corpus-statistics subsets in both languages. For other tasks, where a compact set of supporting evidence is sufficient, we report Recall@5 together with a task-specific complementary retrieval metric (e.g., Precision@5, MRR). Efficiency is measured by end-to-end token usage, including all model input and output tokens. Appendix F provides the complete task-to-metric mapping, metric formulas and details of LLM judge.

3 Experiments

3.1 Experimental Setup

We evaluate all applicable methods on the law, finance, and scientific-research tasks in ANSER-Bench. Experiments are conducted with two strong language-model backbones, **Qwen3.5-122B-A10B** (Qwen Team, 2026) and **DeepSeek-V4.1-Flash** (DeepSeek-AI, 2026).

Table 3: Primary answer scores (↑) for the 14 task settings. D/P/R: descriptive/predictive/prescriptive. Primary metrics: Law D and the finance corpus-statistics subsets use RA@10%; Law P uses relation accuracy; finance filing analysis (D) and forecasting (P) use accuracy on categorical items; Finance-EN R averages the RegBench LLM-judge score and FinAuditing rule accuracy; other R settings use LLM-judge scores; science uses Token-F1. Finance D averages its two subsets equally. Bold marks the best result within a backbone and setting.

	Law						Finance						Science	
	ZH			EN			ZH			EN			EN	
Method	D	P	R	D	P	R	D	P	R	D	P	R	D	P
Qwen3.5-122B-A10B														
Vanilla LLM	0.00	34.00	70.34	0.00	15.79	76.67	23.15	3.95	21.98	33.60	16.67	17.71	27.09	55.77
Sparse RAG	6.03	66.25	65.28	0.00	**58.95**	79.28	21.48	23.68	27.79	17.19	**38.33**	25.61	24.08	61.28
Dense RAG	10.05	62.65	67.41	1.04	0.00	86.14	28.09	27.63	31.19	19.35	20.00	28.28	28.23	**66.24**
DeerFlow	8.00	**72.15**	66.11	15.00	56.00	80.00	36.95	**38.16**	20.97	16.73	25.00	28.57	22.86	56.66
ReAct	4.50	58.23	62.13	0.00	55.10	85.38	43.70	34.21	32.65	**38.56**	25.00	**42.29**	24.46	62.14
Tongyi DR	11.50	70.83	69.02	8.50	42.31	84.79	**51.53**	36.84	**33.25**	33.38	26.67	39.34	30.49	58.64
FinSight	7.50	69.14	**71.23**	0.50	48.28	85.30	40.92	31.58	30.09	16.91	18.33	31.28	28.10	64.15
LawThinker	**24.00**	61.46	70.07	**39.50**	51.81	**87.30**	44.26	35.53	29.60	33.04	20.00	35.66	**34.59**	64.94
DeepSeek-V4.1-Flash														
Vanilla LLM	0.00	47.00	76.61	0.00	53.33	89.68	29.63	44.74	39.88	35.94	48.33	25.24	43.43	60.95
Sparse RAG	13.33	**73.91**	62.57	0.00	59.09	90.61	22.34	22.37	28.71	24.12	**50.00**	20.32	36.92	62.54
Dense RAG	11.00	65.88	75.30	1.00	55.88	92.95	21.80	40.79	39.88	29.69	40.00	31.70	**46.78**	65.24
DeerFlow	9.47	65.06	72.63	21.61	61.90	89.39	42.08	39.47	**43.24**	36.44	31.67	35.12	30.67	**66.97**
ReAct	10.50	67.39	74.92	0.50	54.22	**96.61**	51.91	38.16	40.97	40.66	35.00	**65.86**	34.98	63.82
Tongyi DR	32.00	71.28	**76.64**	46.50	**86.59**	91.84	52.08	38.16	42.36	56.28	40.00	60.74	40.96	62.08
FinSight	12.00	72.63	73.46	1.00	52.56	92.40	45.31	35.53	42.43	33.31	31.67	52.83	44.63	60.50
LawThinker	**41.00**	70.71	72.44	**53.27**	83.16	95.25	**59.99**	**48.68**	41.93	**58.20**	38.33	56.14	34.62	58.31

Baselines. We compare eight representative systems covering **four major paradigms**, with detailed implementation and reproduction settings provided in Appendix G:

(1) Non-retrieval LLM. *Vanilla LLM* directly answers each query using only the language model, without access to the benchmark retrieval corpus.

(2) Retrieval-augmented generation. We implement two standard RAG baselines (Lewis et al., 2020) with different retrieval mechanisms: *Sparse RAG* uses BM25 (Robertson & Zaragoza, 2009), while *Dense RAG* uses Qwen3-Embedding-4B (Zhang et al., 2025).

(3) General-purpose research agents. *ReAct* (Yao et al., 2023), *DeerFlow* (ByteDance, 2025), and *Tongyi DeepResearch* (hereafter *Tongyi DR*) (Tongyi DeepResearch Team et al., 2025) represent agentic approaches that iteratively interact with retrieval or analytical tools and perform multi-step reasoning before producing a final conclusion.

(4) Domain-specialized agents. *FinSight* (Jin et al., 2026) and *LawThinker* (Yang et al., 2026) are originally designed for financial and legal analysis. We evaluate both systems across all benchmark domains to examine whether domain-specific capabilities transfer to other analytical-search tasks.

3.2 Main Results

Overall results. Table 3 reports the designated primary score for each of the 14 task settings, under both backbones. Appendix H reports the compact macro summary, all secondary answer metrics, evidence metrics, efficiency metrics, and repeated-run stability analyses on selected task settings. Across the 14 task settings, LawThinker achieves five best primary scores, Tongyi DeepResearch, ReAct, and DeerFlow each achieve two, Sparse RAG achieves two and Dense RAG achieves one; all maxima use DeepSeek-V4.1-Flash. Current systems still exhibit substantial performance gaps on several task families: the best scores on legal descriptive tasks are only 41.00–53.27, the best Chinese financial prescriptive score is 43.24, and scientific descriptive Token-F1 peaks at 46.78. Based on these results, we further analyze system behavior around three research questions as follows.

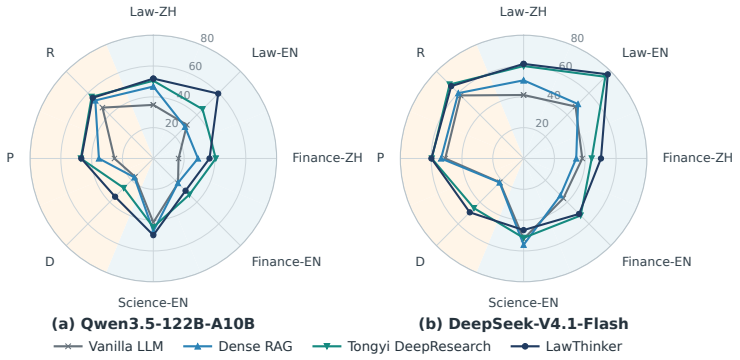


Figure 3: Macro-performance profiles under domain–language aggregates and analytical-need aggregates. Scores are primary-metric macro averages. For readability, we show four representative baselines, covering four baseline paradigms.

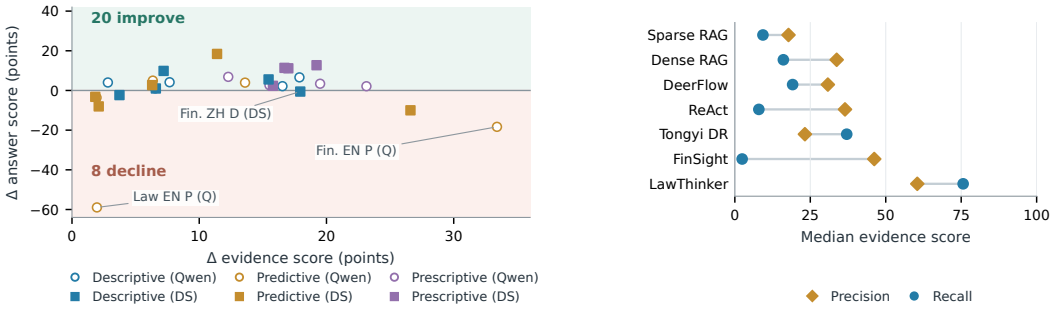
RQ1: How do different methods perform across domains and languages? Figure 3 presents two summaries: five domain–language aggregates and three analytical-need aggregates. Across methods, descriptive analysis has the lowest leading macro score under both Qwen and DeepSeek (35.08/49.42), followed by predictive (49.70/59.84) and prescriptive analysis (56.60/69.59). The lower descriptive scores are consistent with the difficulty of target population aggregation. Across domain–language groups, the leading legal scores are comparatively high (Law-ZH/EN: 51.84/59.54 with Qwen; 61.38/77.23 with DeepSeek), while finance is weaker (Fin-ZH/EN: 40.54/35.28; 50.20/52.34). LawThinker leads both legal groups with either backbone, but the finance leaders vary across Tongyi DR, ReAct, and LawThinker; science is led by LawThinker with Qwen and Dense RAG with DeepSeek.

Domain specialization also transfers selectively. With DeepSeek, LawThinker leads both financial corpus-statistics subsets and Chinese financial prediction, yet ranks seventh/eighth on scientific descriptive/predictive tasks (34.62/58.31). FinSight, despite strong results in some financial and legal settings, scores only 6.00/1.00 RA@10% on Chinese/English corpus statistics. An agent's domain of origin therefore does not determine the reach of its workflow.

Language and backbone also change method rankings. Among the six domain-need pairs with corresponding Chinese and English settings (Law D/P/R and Finance D/P/R), the top-performing method differs between languages in five families under Qwen and four under DeepSeek. For example, with DeepSeek, FinSight outperforms Tongyi DR on Chinese filing analysis (84.62 vs. 73.15) but trails it on English filing analysis (65.63 vs. 76.56). DeepSeek improves 99 of 112 comparable setting scores over Qwen, yet LawThinker's scientific prediction falls from 64.94 to 58.31. Overall, method advantages are sensitive to both language-specific settings and the underlying LLM backbone.

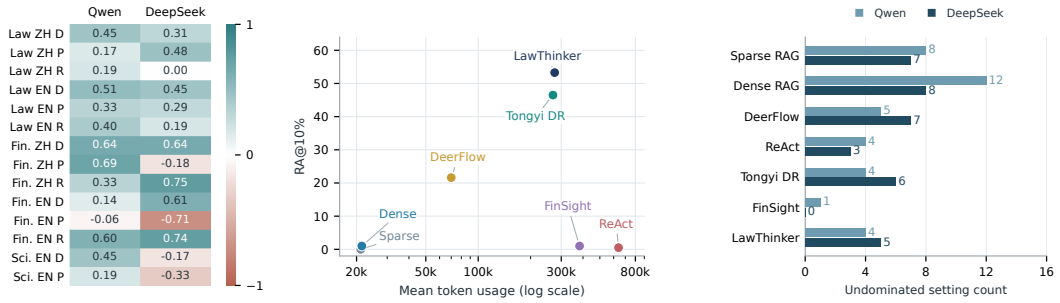
RQ2: Does better evidence retrieval lead to better answers? Dense RAG improves evidence scores over Sparse RAG in all 28 task-setting/backbone pairs. Among these pairs, only 20 answer scores improve (11/14 with Qwen; 9/14 with DeepSeek), while eight decline (Figure 4a). Seven regressions lie outside financial descriptive tasks, spanning legal descriptive/predictive and financial predictive tasks. On Qwen Law-EN-Prediction, set F1 rises by 1.96 points while relation accuracy falls by 58.95; on English financial prediction, Recall@5 rises by 33.40/26.60 with Qwen/DeepSeek, but accuracy falls by 18.33/10.00. This mismatch can arise for at least two reasons. First, better retrieval may identify more relevant evidence units without reliably extracting and using the key numerical values or contextual conditions they contain. Second, stronger evidence acquisition does not eliminate downstream analytical demands such as cross-evidence integration, numerical computation, and relation inference; directly concatenating retrieved evidence can further introduce redundancy and distract generation. Thus, improvements in evidence quality do not automatically translate into improvements in answer quality.

Population-level tasks (as defined in Section 2.3) further expose a coverage gap (Figure 4b). Across the 12 corresponding settings, FinSight achieves a median set precision of 46.19 but only 2.42 recall, indicating that it often retrieves relevant evidence while covering only a small fraction of the target population. In contrast, LawThinker attains substantially higher precision and recall (60.45/75.65).



(a) Dense vs. Sparse: Answer and Evidence Gains **(b) Population-level evidence coverage**

Figure 4: Evidence acquisition and answer quality. (a) Dense-over-Sparse gains across all 28 task-setting/backbone pairs; differences are computed as $\Delta s = s_{\text{dense}} - s_{\text{sparse}}$ for both answer and evidence scores. (b) Median set precision and recall across all 12 population-level evaluation subsets: law descriptive/predictive and finance statistics, in both languages and backbones.



(a) Compute-Performance Correlation Spearman ρ **(b) Efficiency-Performance Trade-off** Law-EN-Descriptive, DeepSeek **(c) Pareto-Optimal Frequency**

Figure 5: Token-usage trade-offs. (a) Spearman's ρ between mean token usage and primary answer score across methods for each setting and backbone. (b) Token usage versus RA@10% on Law-EN descriptive tasks with DeepSeek-V4.1-Flash. (c) Number of task settings in which each method lies on the Pareto frontier. A method is Pareto-optimal if no other method achieves better answer quality with the same or lower token cost, or lower token cost with the same or better answer quality.

These results highlight that analytical search requires not only relevant evidence, but also sufficient population coverage to support reliable aggregation and association analysis.

RQ3: Does more computation yield better analytical performance? We compute the Spearman correlation between token usage and answer quality separately within each task setting and backbone, yielding 28 correlations (Figure 5a). Correlations are positive in 22/28 cases, with medians of 0.37/0.30 for Qwen/DeepSeek. Chinese financial descriptive tasks show $\rho = 0.64/0.64$, whereas English financial prediction has $\rho = -0.06/-0.71$; both scientific tasks are also negatively correlated under DeepSeek. These negative cases show that additional computation is not uniformly productive: higher token usage can coincide with worse answer quality. Possible mechanisms behind this include redundant retrieval, noisy context, or ineffective additional reasoning.

To deeply assess whether the additional cost is worthwhile, we further examine Pareto efficiency. A method is *dominated* if another method achieves at least as good answer quality with no greater token cost, and is strictly better on either quality or cost; the remaining methods form the *Pareto frontier*. For example, on DeepSeek Law-EN-D tasks, LawThinker achieves 53.27 RA@10% with 275k tokens, whereas ReAct uses 640k tokens for 0.50 and is therefore dominated (Figure 5b). Across task settings, Dense RAG appears on the frontier in 20/28 settings, compared with 1/28 for FinSight. More broadly, the highest-token method is dominated in 26/28 task-setting/backbone pairs, indicating that simply spending more computation rarely yields the best cost–quality trade-off.

Beyond the quantitative analyses above, we also conduct error analyses on representative cases to examine common failure modes in analytical-search pipelines. Details are provided in Appendix I.

4 Related Work

From Information Retrieval to Analytical Search. Traditional information retrieval is centered on ranking documents by relevance, assuming that users inspect retrieved results and complete subsequent analysis themselves. Early work on complex and interactive question answering already challenged this assumption: HITIQA supported analysts in iteratively exploring unstructured collections to resolve complex analytical problems (Small & Strzalkowski, 2009), while exploratory search emphasized search as a process of learning, investigation, and sensemaking rather than simple fact finding (Marchionini, 2006). Later, open-domain and multi-hop QA shifted more evidence composition into the system, and retrieval-augmented generation further connected retrieval with downstream answer generation (Yang et al., 2018; Lewis et al., 2020). Systems such as WebGPT and ReAct subsequently introduced iterative browsing, evidence collection, and tool-augmented reasoning (Nakano et al., 2022; Yao et al., 2023), moving search systems closer to automated problem solving. *Analytical search* consolidates these developments into an end-to-end paradigm centered on evidence acquisition, analytical operations, and verifiable conclusions (Tu et al., 2026), which provides the conceptual basis for ANSER-Bench.

Related Benchmarks. Existing benchmarks cover different components of analytical search. **Retrieval-oriented** benchmarks such as BEIR and MultiConIR evaluate general retrieval and complex multi-condition matching (Thakur et al., 2021; Lu et al., 2025). **Reasoning-oriented** benchmarks such as HotpotQA and MultiHop-RAG require evidence composition across multiple documents (Yang et al., 2018; Tang & Yang, 2024), while STARQA and GlobalQA extend analysis to structured databases and corpus-level operations such as filtering, counting, sorting, and aggregation (Maddela et al., 2025; Luo et al., 2025). More recent **end-to-end research benchmarks**, including DeepResearch Bench and FinSearchComp, further evaluate iterative information gathering, synthesis, and professional search (Du et al., 2026; Hu et al., 2026). Overall, existing benchmarks cover complementary pieces of the analytical-search problem, but differ substantially in retrieval scope, evidence form, and analytical objective.

Methods for Analytical Search. Methods relevant to analytical search can be organized into three lines. **Retrieval and RAG** methods use sparse or dense retrieval to acquire external evidence and condition generation on retrieved context (Karpukhin et al., 2020; Lewis et al., 2020). **Tool-augmented and agentic methods** extend this pipeline with query decomposition, iterative retrieval, computation, and external actions, enabling longer-horizon problem solving (Yao et al., 2023; Tongyi DeepResearch Team et al., 2025). More recently, **domain-specialized agents** incorporate professional workflows and domain-specific reasoning, exemplified by FinSight in finance and LawThinker in law (Jin et al., 2026; Yang et al., 2026). Together, these lines reflect a broader shift from isolated retrieval or generation toward integrated systems that coordinate evidence acquisition, reasoning, and analysis.

5 Conclusion

We introduce **ANSER-Bench**, a benchmark for analytical search comprising 2,235 Chinese and English tasks across law, finance, and scientific research, with explicit evaluation of both analytical conclusions and their supporting evidence. Experiments across eight methods and two LLM backbones show that performance remains highly task-dependent, and that neither stronger retrieval nor greater computation consistently translates into better final answers. Reliable analytical search therefore requires systems to jointly acquire sufficient evidence, apply task-appropriate analytical operations, and produce verifiable conclusions. We hope ANSER-Bench provides a common testbed for developing and diagnosing such systems.

At the same time, the benchmark has several limitations. Its current coverage is restricted to three domains and two languages, and the use of fixed corpora abstracts away some open-ended source discovery and dynamic tool use in real-world analytical workflows. In addition, although evidence and final answers are evaluated separately, failures may still arise from interactions among retrieval, structured querying, aggregation, computation, reasoning, and generation, making precise error attribution difficult. Extending the benchmark to broader settings and developing finer-grained component-level diagnostics remain important directions for future work.

AI USE STATEMENT

We used generative AI tools to assist with literature retrieval and discovery, research ideation and execution, drafting and polishing parts of the manuscript, and limited generation of synthetic benchmark instances. During benchmark construction, LLMs assisted with extracting structured fields from source documents, expanding and rewriting candidate queries, assessing candidate evidence, and drafting or cross-checking some reference answers. Human annotators reviewed all queries, independently verified reference conclusions, adjudicated disagreements in evidence assessment, and manually audited a sample of evidence annotations. In the experiments, LLMs served as the backbones of the evaluated systems; an LLM judge scored specified answer types using the rubrics described in Appendix F and was validated against human ratings. The authors reviewed the AI-assisted work and take responsibility for the final benchmark, methods, results, references, and manuscript.

ETHICS AND REPRODUCIBILITY

ANSER-Bench is intended for research on professional information analysis. Corpus access does not establish redistribution rights or suitability for legal, financial, or clinical decision-making. During review, our anonymous repository provides the benchmark queries, reference answers and evidence annotations, evaluation code and judge prompts, and small-scale corpus samples. Appendix J details the release plan, versioning and documentation, licensing, and intended use.

REFERENCES

Administrative Office of the U.S. Courts. Civil cover sheet (form JS 44), n.d. URL https://www.uscourts.gov/forms-rules/forms/civil-cover-sheet. Accessed September 2026.

AKShare. Financial statements from eastmoney: balance sheet, income statement, and cash-flow statement interfaces, n.d. URL https://akshare.akfamily.xyz/data/stock/stock.html#id195. Accessed 25 September 2026.

ByteDance. DeerFlow: Deep exploration and efficient research flow. GitHub repository, 2025. URL https://github.com/bytedance/deer-flow. Accessed 18 September 2026.

China Securities Regulatory Commission. Standards for the contents and formats of information disclosure by companies offering securities to the public no. 2: Annual report contents and format (2021 revision), 2021. URL https://www.csrc.gov.cn/csrc/c101864/c6df1268b5b294448bdec7e010d880a01/content.shtml.

Jonathan J. Deeks, Julian P. T. Higgins, Douglas G. Altman, Joanne E. McKenzie, and Areti A. Veroniki. Chapter 10: Analysing data and undertaking meta-analyses. In *Cochrane Handbook for Systematic Reviews of Interventions, Version 6.5*. Cochrane, 2024. URL https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-10.

DeepSeek-AI. DeepSeek-V4.1-Flash: Pushing the limits of kv cache compression. *arXiv preprint arXiv:2609.19969*, 2026. doi: 10.48550/arXiv.2609.19969. URL https://arxiv.org/abs/2609.19969.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Licheng Zhang, Xiaorui Wang, and Zhendong Mao. DeepResearch Bench: A Comprehensive Benchmark for Deep Research Agents. In *International Conference on Learning Representations*, pp. 42414–42448, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/465f22be10e07b301c6ed58f0472f704-Abstract-Conference.html.

Free Law Project. CourtListener: Bulk legal data, n.d. URL https://wiki.free.law/c/courtlistener/help/api/bulk-data/bulk-legal-data. Accessed September 2026.

Abe Bohan Hou, Orion Weller, Guanghui Qin, Eugene Yang, Dawn Lawrie, Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. CLERC: A dataset for U.S. legal case retrieval and retrieval-augmented analysis generation. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 7913–7928, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.441. URL https://aclanthology.org/2025.findings-naacl.441/.

Chuxuan Hu, Maxwell Yang, James Weiland, Yeji Lim, Suhas Palawala, and Daniel Kang. DRAMA: Unifying data retrieval and analysis for open-domain analytic queries. *Proceedings of the ACM on Management of Data*, 3(6):1–28, 2025. doi: 10.1145/3769781. URL https://doi.org/10.1145/3769781.

Liang Hu, Jianpeng Jiao, Jiashuo Liu, Dongyuan Mutu, Yanle Ren, Zhoufutu Wen, Kaiyuan Zhang, Xuanliang Zhang, Xiang Gao, Tianci He, Fei Hu, Yali Liao, Zaiyuan Wang, Jingkai Liu, Daibin Sun, Ziqing Zeng, Zhiyuan Zeng, Chenghao Yang, Qianyu Yang, Mingren Yin, Xinying Feng, Ge Zhang, Xinyi Zhang, Xiying Zhao, Zhenwei Zhu, Hongseok Namkoong, and Wenhao Huang. FinSearchComp: Towards a realistic, expert-level evaluation of financial search and reasoning. In *International Conference on Learning Representations*, pp. 46402–46428, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/4d42358702dff82e1436550a05ade260-Abstract-Conference.html.

Jiajie Jin, Yuyao Zhang, Yimeng Xu, Hongjin Qian, Yutao Zhu, and Zhicheng Dou. FinSight: Towards real-world financial deep research. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5868–5894, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.265. URL https://aclanthology.org/2026.acl-long.265/.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550/.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

Xuan Lu, Sifan Liu, Bochao Yin, Yongqi Li, Xinghao Chen, Hui Su, Yaohui Jin, Wenjun Zeng, and Xiaoyu Shen. MultiConIR: Towards multi-condition information retrieval. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 13471–13494, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.726. URL https://aclanthology.org/2025.findings-emnlp.726/.

Qi Luo, Xiaonan Li, Tingshuo Fan, Xinchi Chen, and Xipeng Qiu. Towards Global Retrieval Augmented Generation: A Benchmark for Corpus-Level Reasoning. *arXiv preprint arXiv:2510.26205*, 2025. doi: 10.48550/arXiv.2510.26205. URL https://arxiv.org/abs/2510.26205.

Mounica Maddela, Lingjue Xie, Daniel Preotiuc-Pietro, and Mausam. STARQA: A question answering dataset for complex analytical reasoning over structured databases. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 34487–34499,

Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1749. URL https://aclanthology.org/2025.emnlp-main.1749/.

Gary Marchionini. Exploratory search: From finding to understanding. *Communications of the ACM*, 49(4):41–46, 2006. doi: 10.1145/1121949.1121979.

Joanne E. McKenzie, Sue E. Brennan, Rebecca E. Ryan, Hilary J. Thomson, and Renea V. Johnston. Chapter 9: Summarizing study characteristics and preparing for synthesis. In *Cochrane Handbook for Systematic Reviews of Interventions, Version 6.5*. Cochrane, 2024. URL https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-09.

Donald Metzler, Yi Tay, Dara Bahri, and Marc Najork. Rethinking search: Making domain experts out of dilettantes. *ACM SIGIR Forum*, 55(1):13:1–13:27, 2021. doi: 10.1145/3476415.3476428.

Ministry of Finance of the People's Republic of China. Accounting standard for business enterprises no. 31: Statement of cash flows, 2006. URL https://kjs.mof.gov.cn/zt/kjzzss/kuaijizhunzeshishi/200806/t20080618_46250.htm.

Ministry of Finance of the People's Republic of China. Accounting standard for business enterprises no. 30: Presentation of financial statements (2014 revision), 2014. URL https://kjs.mof.gov.cn/zt/kjzzss/kuaijizhunzeshishi/200806/t20080618_46218.htm.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. WebGPT: Browser-assisted question-answering with human feedback, 2022. URL https://arxiv.org/abs/2112.09332.

Jane A. Ou and Stephen H. Penman. Financial statement analysis and the prediction of stock returns. *Journal of Accounting and Economics*, 11(4):295–329, 1989. ISSN 0165-4101. doi: 10.1016/0165-4101(89)90017-7. URL https://www.sciencedirect.com/science/article/pii/0165410189900177.

Matthew J. Page et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372:n71, 2021. doi: 10.1136/bmj.n71. URL https://www.bmj.com/content/372/bmj.n71.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

RegBench Team. Regbench: Source-grounded benchmarks for regulatory cross-reference reasoning. Project repository, 2026. URL https://github.com/regbench/regbench.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, April 2009. ISSN 1554-0669. doi: 10.1561/1500000019. URL https://doi.org/10.1561/1500000019.

SALI Alliance. The legal matter specification standard, n.d. URL https://sali.org/explore-the-standard/. Accessed September 2026.

Shanghai Stock Exchange. Listed-company periodic reports, n.d. URL https://www.sse.com.cn/disclosure/listedinfo/regular/. Accessed 25 September 2026.

Sharon Small and Tomek Strzalkowski. HITIQA: High-quality intelligence through interactive question answering. *Natural Language Engineering*, 15(1):31–54, 2009. doi: 10.1017/S1351324908004890.

Software Repository for Accounting and Finance, University of Notre Dame. SEC/EDGAR data: Cleaned and raw 10-X files, n.d. URL https://sraf.nd.edu/sec-edgar-data/. Accessed 25 September 2026.

Weihang Su, Yiran Hu, Anzhe Xie, Qingyao Ai, Quezi Bing, Ning Zheng, Yun Liu, Weixing Shen, and Yiqun Liu. STARD: A chinese statute retrieval dataset derived from real-life queries by non-professionals. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10658–10671. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.625. URL https://aclanthology.org/2024.findings-emnlp.625/.

Yixuan Tang and Yi Yang. MultiHop-RAG: Benchmarking retrieval-augmented generation for multi-hop queries. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=t4eB3zYWBK.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In J. Vanschoren and S. Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/65b9eea6e1cc6bb9f0cd2a47751a186f-Paper-round2.pdf.

Tongyi DeepResearch Team, Baixuan Li, Bo Zhang, Dingchu Zhang, Fei Huang, Guangyu Li, Guoxin Chen, Huifeng Yin, Jialong Wu, Jingren Zhou, Kuan Li, Liangcai Su, Litu Ou, Liwen Zhang, Pengjun Xie, Rui Ye, Wenbiao Yin, Xinmiao Yu, Xinyu Wang, Xixi Wu, Xuanzhong Chen, Yida Zhao, Zhen Zhang, Zhengwei Tao, Zhongwang Zhang, Zile Qiao, Chenxi Wang, Donglei Yu, Gang Fu, Haiyang Shen, Jiayin Yang, Jun Lin, Junkai Zhang, Kui Zeng, Li Yang, Hailong Yin, Maojia Song, Ming Yan, Minpeng Liao, Peng Xia, Qian Xiao, Rui Min, Ruixue Ding, Runnan Fang, Shaowei Chen, Shen Huang, Shihang Wang, Shihao Cai, Weizhou Shen, Xiaobin Wang, Xin Guan, Xinyu Geng, Yingcheng Shi, Yuning Wu, Zhuo Chen, Zijian Li, and Yong Jiang. Tongyi DeepResearch Technical Report. *arXiv preprint arXiv:2510.24701*, 2025. doi: 10.48550/arXiv.2510.24701. URL https://arxiv.org/abs/2510.24701.

Yiteng Tu, Shuo Miao, Weihang Su, Yiqun Liu, and Qingyao Ai. Analytical search. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2747–2757. Association for Computing Machinery, 2026. doi: 10.1145/3805712.3808536. URL https://doi.org/10.1145/3805712.3808536.

U.S. Securities and Exchange Commission. Commission guidance regarding management's discussion and analysis of financial condition and results of operations, 2003. URL https://www.sec.gov/rules-regulations/2003/12/commission-guidance-regarding-managements-discussion-analysis-financial-condition-results-operations. Release Nos. 33-8350; 34-48960.

U.S. Securities and Exchange Commission. Financial statement data sets, n.d. URL https://www.sec.gov/data-research/sec-markets-data/financial-statement-data-sets. Accessed 25 September 2026.

Yan Wang, Keyi Wang, Shanshan Yang, Jaisal Patel, Jeff Zhao, Fengran Mo, Xueqing Peng, Lingfei Qian, Yankai Chen, Víctor Gutiérrez-Basulto, Jimin Huang, Guojun Xiong, Xiao-Yang Liu, and Jian-Yun Nie. FinAuditing: A financial taxonomy-structured multi-document benchmark for evaluating LLMs. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3456–3463. Association for Computing Machinery, 2026. doi: 10.1145/3805712.3808578. URL https://doi.org/10.1145/3805712.3808578.

Anzhe Xie, Weihang Su, Yujia Zhou, Yiqun Liu, Min Zhang, and Qingyao Ai. MetaSyn: A benchmark for LLM agents on meta-analysis articles from nature portfolio, 2026. URL https://arxiv.org/abs/2606.17041.

Xinyu Yang, Chenlong Deng, Tongyu Wen, Binyu Xie, and Zhicheng Dou. LawThinker: A Deep Research Legal Agent in Dynamic Environments. *arXiv preprint arXiv:2602.12056*, 2026. doi: 10.48550/arXiv.2602.12056. URL https://arxiv.org/abs/2602.12056.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question

answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models, 2025. URL https://arxiv.org/abs/2506.05176.

APPENDIX OVERVIEW

- Appendix A: Detailed Task Formulation
- Appendix B: Benchmark Overview
- Appendix C: Legal Benchmark Construction
- Appendix D: Financial Benchmark Construction
- Appendix E: Scientific Benchmark Details
- Appendix F: Evaluation Details
- Appendix G: Baseline Implementations and Reproduction Settings
- Appendix H: Full Experimental Results
- Appendix I: Error Analysis
- Appendix J: Release and Reproducibility

A DETAILED TASK FORMULATION

Section 2.1 represents an analytical-search instance as $x = (q, \mathcal{C}, E^*, y^*, m)$. This section specifies the analytical objectives shared across task domains. The accessible corpus $\mathcal{C} = \{\mathcal{D}^u, \mathcal{D}^s\}$ contains unstructured sources and structured representations; the same underlying source may have both. Metadata m specifies the task's scope and required answer form. A system uses q and $\mathcal{C}$ to produce an answer $\hat{y}$ and supporting evidence $\hat{E}$, while y^* and E^* remain reference annotations for evaluation.

A.1 SHARED ANALYTICAL INTERFACE

Let $\mathcal{O}_q(\mathcal{C})$ be the logical units over which q asks the system to reason, and let Φ_q be its selection or eligibility conditions. A unit may be represented by multiple records or passages. Conditions can be checked against structured fields, unstructured content, or their combination. The formulations below describe the target operation; they do not prescribe how a system must implement it or which intermediate steps it must expose.

A.2 DESCRIPTIVE TASKS (D)

Descriptive tasks determine what is present in observed evidence. They select units satisfying the query conditions and apply a descriptive operator:

$$S_q = \big\{o \in \mathcal{O}_q(\mathcal{C}) : \textstyle\bigwedge_{\phi \in \Phi_q} \phi(o; \mathcal{C}) = 1\big\}, \qquad y^*_{\mathrm{D}} = g_q(S_q, \mathcal{C}). \tag{2}$$

Here S_q is the qualifying set, and g_q may extract a value, count, aggregate, compare, rank, or summarize properties of that set. The answer may therefore be a number, category, ordering, or concise synthesis rather than a retrieved passage. When q asks for a population-level result, supporting evidence must cover the qualifying units sufficiently to justify the aggregation; when q asks for a local fact or pattern, a smaller decisive evidence set may suffice. The difficulty lies in applying all conditions consistently and acquiring enough evidence for the requested operation.

A.3 PREDICTIVE TASKS (P)

Predictive tasks infer a target that is not directly available as a single observed fact. The target may be a relationship or effect inferred from observations, or an outcome beyond an information cut-off. Let S_q be the eligible set defined as in Eq. (2), and let $\mathbf{z}_q = \mathrm{Extract}_q(S_q, \mathcal{C})$ collect the relevant measurements, attributes, outcomes, and comparison information. If q specifies slices Δ_q, such as groups or time periods, $\mathbf{z}_q$ may include their summaries $\{g_{q,\delta}(S_{q,\delta}, \mathcal{C})\}_{\delta \in \Delta_q}$. For a target defined over observed evidence,

$$y^*_{\mathrm{P},obs} = h_q(\mathbf{z}_q), \tag{3}$$

where h_q estimates or synthesizes the requested relationship and maps it to the required answer form. Such a result may express direction, strength, or uncertainty; an observed association alone does not establish causation.

For a temporally prospective target, q defines an information cut-off t_0 and a later target time t_1. Write $\mathcal{C}_{\leq t_0}$ for the portion of the corpus available by t_0. The system's answer must be based on that portion, while the reference can be determined from the later realized target z_{t_1}:

$$\hat{y}_{\mathrm{P},fut} = f_q(\mathcal{C}_{\leq t_0}), \qquad y^*_{\mathrm{P},fut} = \mathrm{label}_q(z_{t_1}), \qquad t_1 > t_0. \tag{4}$$

The function f_q denotes the evaluated system's analysis, not a prescribed forecasting model. The later target supplies supervision and is outside the permitted input. Across both forms, the supporting evidence must match the population, comparison, and information boundary specified by q.

A.4 Prescriptive tasks (R)

Prescriptive tasks determine what conclusion or action follows when observed facts are assessed against applicable criteria, constraints, and any stated objective. Let F_q be the relevant facts extracted from $\mathcal{C}$, let $K_q \subseteq \mathcal{D}^u \cup \mathcal{D}^s$ denote the applicable criteria as represented in the corpus, and let U_q be an optional objective specified by q. A criterion may be expressed schematically as

$$k_j: \quad \mathrm{Cond}_j(F_q) \ \wedge \ \neg\, \mathrm{Exc}_j(F_q) \implies \mathrm{Conseq}_j(F_q), \tag{5}$$

where Cond_j gives the conditions for application, Exc_j an exception if one exists, and Conseq_j the resulting judgment, obligation, or value. This implication is an abstraction: some criteria require analogical comparison rather than literal rule matching. The target is

$$y^*_{\mathrm{R}} = J_q(F_q, K_q; U_q), \tag{6}$$

where J_q selects relevant criteria, tests their conditions and exceptions, and applies them to the facts and any stated objective. The answer can be a judgment, required action, calculated value, or an explicit statement that the available facts do not determine a conclusion. Its support must connect the material facts and applicable criteria to the result.

Table 4 summarizes the three objectives. They are distinguished by the conclusion sought, although solving one task may involve operations associated with another: an inference may require descriptive aggregation, and a prescriptive judgment may require a calculation.

Table 4: Analytical needs, core operations, and expected outputs. P includes historical association and evidence synthesis as well as future forecasting.

Need	Core operation	Output and supporting evidence
D: Descriptive	Select eligible units and extract, count, aggregate, compare, rank, or synthesize observed evidence.	A fact, count, statistic, category, or concise pattern; a qualifying population or decisive sources.
P: Predictive	Estimate a relation or effect from eligible observations, or infer a later outcome from pre-cut-off evidence.	An association/effect conclusion with requested statistics, or a forecast label; cohort, studies, or historical records.
R: pRescriptive	Match facts to applicable rules or precedents, check conditions and exceptions, then apply them.	A source-grounded judgment, obligation, or rule/value answer; relevant facts and normative support.

B BENCHMARK OVERVIEW

Table 5 summarizes the benchmark settings across domains, languages, and analytical task types, including corpus scale, answer form, evidence form, and query count.

Table 5: ANSER-Bench task settings. ZH/EN: Chinese/English; D/P/R: descriptive/predictive/prescriptive.

Setting	Subtask	Corpus	Answer form	Evidence form	#Query
Law-ZH-D	Case statistics	18,423 judicial decisions; 292,594 case features	Conditioned case counts; integer	Qualifying cases	200
Law-ZH-P	Factor–outcome association		Association direction and requested statistics; JSON object	Eligible cases	100
Law-ZH-R	Statutory rule application	55,396 statutory and judicial-interpretation provisions	Legal judgment, basis, and rationale; short text	Applicable statutory provisions	100
Law-EN-D	Case statistics	12,708 cases; 13,508 opinions; 414,959 case features; 67,260 documents	Conditioned case counts; integer	Qualifying opinion clusters	200
Law-EN-P	Factor–outcome association		Association direction and requested statistics; JSON object	Eligible opinion clusters	95
Law-EN-R	Precedent application	5,000 precedent documents	Governing rule and application; short text	Relevant precedent documents	57
Finance ZH D	Filing analysis	27,089 financial reports	Facts, comparisons, screening, or synthesis; numeric (12), categorical (108), or short text (80)	Relevant financial reports	300
	Corpus statistics		Target population statistics; numeric (100)	Reports mapped to distinct companies	
Finance ZH P	Financial forecasting		Future values, ranges, trends, or scenarios; numeric (24), categorical (76), or short text (100)	Historical financial reports	200
Finance ZH R	Disclosure compliance	216 financial regulatory rules	Disclosure-compliance judgment and rationale; short text	Applicable laws and articles	100
Finance EN D	Filing analysis	21,220 filings; 39,139 report sections	Facts, comparisons, screening, or synthesis; numeric (47), categorical (64), or short text (89)	Relevant filing accessions	300
	Corpus statistics		Target population statistics; numeric (100)	Company identifiers linked to filings	
Finance EN P	Financial forecasting		Future values, categories, or narratives; numeric (60), categorical (60), or short text (80)	Historical filing accessions	200
Finance EN R	Capital-rule assessment	3,329 regulation paragraphs	Capital-rule conclusion and application; short text (200)	Relevant regulation paragraphs	300
	Filing-fact audit	137 DQC audit rules; 5,877 filing facts	Applied audit rule and expected value; JSON object (100)	Relevant filing facts	
Science EN D	Pattern synthesis	140,585 scientific articles	Main factors, dominant category, or study pattern; short text	Eligible primary articles	20
Science EN P	Effect/association synthesis		Effect or association direction, or diagnostic performance; short text	Eligible primary articles	63
Total					**2,235**

C LEGAL BENCHMARK CONSTRUCTION

C.1 DATA SOURCES AND CORPUS CONSTRUCTION

C.1.1 CHINESE LEGAL CORPUS

Judgment corpus annotation. The Chinese case corpus was collected from publicly available decisions on China Judgments Online (`https://wenshu.court.gov.cn/`), covering 2019–2024. The corpus contains 18,423 cases, including 15,809 civil (85.8%), 2,471 criminal (13.4%), and 143 administrative (0.8%) cases, spanning 91, 64, and five causes of action, respectively. This distribution broadly reflects the case composition encountered in practical judicial settings. Figure 6 summarizes the category distribution and major causes of action.

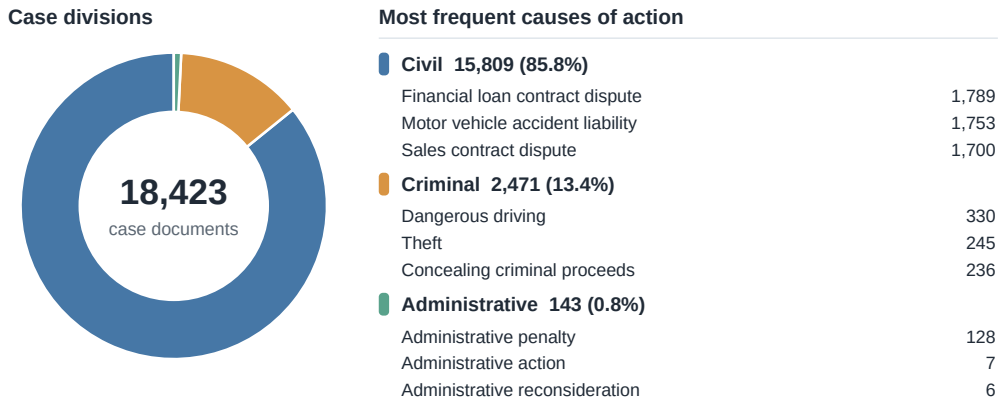


Figure 6: Composition of the final Chinese judgment corpus. Division totals count distinct cases. Leading causes are counted from all labels in each case's `anyou` field, so a multi-cause judgment may contribute to more than one cause count.

Cleaning and normalization. The source decisions combine heterogeneous layouts with broken line wraps, repeated headers and footers, redundant whitespace, garbled characters, and occasional repeated passages. Preprocessing standardizes text encoding and spacing, rejoins disrupted lines, and filters identifiable duplication while preserving titles, paragraph boundaries, case numbers, and numbered legal references needed for later evidence tracing. Cause names and stable metadata are normalized before indexing. The exported database stores one `cases` row per case number and retains the cleaned full text, which avoids counting multiple feature rows as multiple decisions.

Field design and extraction. The representation separates case-wide metadata from facts that vary by cause of action. The former identify the document and support grouping by case, court, procedure, cause, and decision date. The latter, whose fields are defined by legal experts for each cause of action, encode parties, conduct, amounts, circumstances, and outcomes needed to identify sets of cases satisfying specified analytical conditions. Table 6 groups example fields by the meaning of their values. Categories, measurements, free text, and event dates support different comparison and filtering needs.

Table 6: Field families in the final Chinese judgment corpus. The examples are stored column names or feature keys; types describe their analytical meaning rather than their SQLite storage type.

Field family	Function	Examples in the stored corpus
General metadata	Identify the decision and its procedural context.	`anhao` (case number) `anyou` (cause of action) `caipanriqi` (decision date).
List-valued	Represent multiple parties, places, events, or objects in one case.	`beigao_name_list` (defendants) `daoqiewupinleixing_list` (types of stolen property).
Boolean	Record whether a specified factual condition holds.	`has_danbaoren` (guarantor) `fangwu_is_kongzhi` (vacant property) `zaochengjiaotongshigu` (traffic accident).
Categorical	Select a value from a finite set of outcomes or attributes.	`outcome` (judgment result) `daikuanqixian` (loan-term category).
Numeric	Support thresholds, ranges, and aggregation over measurable values.	`jiujinghanliang` (blood alcohol) `daoqieshue` (amount stolen) `wuye_fangwu_mianji` (property area).
Free text	Preserve open-ended descriptions that do not fit fixed labels.	`jiaotongshiguluduan` (accident location) `fangwu_zuoluo_dizhi` (property address).
Date-valued	Locate an underlying event in time, apart from the judgment date.	`gongshang_riqi` (work-injury date) `anjianfashengshijian` (incident time).

Stable metadata, such as case numbers, causes of action, courts, and judgment dates, are obtained directly from the original web records. Other case facts are annotated through a combination of rule-based and LLM-based extraction. Rule-based extraction is used mainly for Boolean and numeric fields, where explicit patterns or thresholds can be reliably identified, such as the amount of stolen property. For the remaining fields, we use Qwen3-32B with its reasoning mode enabled to perform schema-guided structured annotation from the judgment text, mapping extracted facts to the predefined field definitions. All extracted values are normalized to the designed schema and checked against the source decisions during construction. The final judgment corpus contains 174 distinct populated `feature_key`s and 292,594 feature rows linked to 17,341 of the 18,423 cases.

Stored schema. Table 7 summarizes the final case-corpus representation. General metadata and judgment text are stored in `cases`, while typed case-specific features are stored in the many-to-one `case_features` table.

Table 7: Stored schema for the Chinese legal case corpus.

Table	Main fields	Role
`cases`	`anhao`, `anyou`, `wenshuleixing`, `fayuanmingcheng`, `caipanriqi`, `shenpanchengxu`, `wenshu_content`	One row per case number; retains normalized metadata and the full judgment text.
`case_features`	`anhao`, `feature_key`, `feature_value`	Many-to-one structured case facts and outcomes; values may be stored as text or JSON-encoded lists.

Statute corpus annotation. The prescriptive retrieval corpus is adapted from the STARD statute collection (Su et al., 2024), which contains article-level Chinese laws, regulations, and judicial interpretations collected from official sources. As STARD was released in 2024, some legal texts have since been revised. For example, the Anti-Money Laundering Law was revised in November 2024 and took effect in January 2025. We therefore performed an incremental update to the original

statute collection, resulting in a final corpus of 55,396 provision records covering 1,258 named legal texts.

Conversion to the benchmark schema is direct: retain the article text and source identifier, separate the law title and article index into `law_name` and `item`, and write one JSONL record per provision with `id` and `content`.

C.1.2 ENGLISH LEGAL CORPUS

Case corpus annotation. We construct the English case-law corpus from the Free Law Project's CourtListener bulk data (Free Law Project, n.d.), available through `https://www.courtlistener.com/`. We first collected an initial pool of 20,000 cases from 440 courts, with decisions dated 1970–2025 and 94,314 retrieval chunks of up to approximately 800 tokens. CourtListener's opinion clusters provide the case-level grouping; a case may have multiple associated opinion texts. Following screening for usable opinion text and reliable structured annotations, the final database contains 12,708 cases from 414 courts, 13,508 opinion records, and 67,260 retrieval chunks.

Cleaning and normalization. The source includes different opinion types and text representations. We select a usable preferred text for each retained opinion, normalize it for retrieval, and keep its source type, text length, digest, and quality metrics alongside the text. Each opinion is segmented into searchable passages of at most 800 tokens, with both opinion and case identifiers preserved so a retrieved passage can be traced back to its source.

Field design and extraction. The annotation scheme has four linked layers (Table 8). It is informed by CourtListener metadata, the SALI Legal Matter Specification Standard (SALI Alliance, n.d.), and the U.S. Courts' Civil Cover Sheet and Nature of Suit categories (Administrative Office of the U.S. Courts, n.d.). Broad practice areas organize the corpus, matter types specify the kind of dispute, legal issues identify adjudicated questions that may recur across matters, and typed matter features capture the facts and outcomes needed for analytical filtering. The local registry defines 23 practice areas, 120 matter types, 47 legal issues, and 246 feature keys.

Table 8: The four annotation layers in the English legal corpus. Examples are identifiers present in the final annotation registry.

Layer	Analytical role	Example labels or fields
Practice Area	Broad legal domain of the present case.	`administrative_law`; `civil_rights_law`.
Matter Type	Specific dispute or proceeding within a practice area.	`apa_review` (review of agency action); `section_1983_claim` (civil-rights claim under 42 U.S.C. § 1983).
Legal Issue	Question resolved by the court, potentially shared across matter types.	`claim_preclusion`; `harmless_error`.
Matter Type Features	Typed case facts, actors, quantities, and decisions within a matter.	`government_actor_type`; `amount_awarded_usd`; `current_court_disposition`.

Annotation proceeds from rule-based candidate generation to semantic verification, deterministic validation, and targeted review. Rules propose labels from case names, cited statutes, court metadata, and opinion text. GPT-5.4-mini then checks whether a candidate describes the *present* case, rather than a cited case, legal definition, procedural history, hypothetical, or claim the court did not adjudicate. Programmatic validators based on predefined rules and schema constraints enforce closed vocabularies and check evidence scope, actor and target, polarity, units, and consistency between layers. GPT-5.6 Sol re-check failed, low-confidence, and conflicting records; persistent conflicts in a small subset receive manual review. Feature values distinguish `present`, `absent`, `not_mentioned`, `not_applicable`, and `uncertain`, rather than collapsing an unobserved fact into a negative label.

Stored schema. Table 9 summarizes the final case-level, opinion-level, and feature-level representation. A `case_id` joins cases to opinions, features, and retrieval chunks; `cluster_id` retains the

CourtListener opinion-cluster identifier. The registry stores the allowed annotation vocabulary and field definitions used during validation.

Table 9: Stored schema for the English legal case corpus.

Table	Main fields	Role
case_record	case_id, cluster_id, court_id, decision_date, practice_areas_json, matter_types_json, legal_issues_json	One row per retained case, with CourtListener metadata and case-level annotations.
opinion_record	opinion_id, case_id, opinion_type, preferred_text, quality_metrics_json	Opinion text and quality information, linked to its case.
case_feature	case_id, matter_type, feature_key, value_status, value_json, value_number, unit, confidence, evidence_json	Matter-specific structured facts and outcomes.
search_document	document_id, case_id, opinion_id, chunk_text, chunk_token_count	Retrievable opinion passages with source links.

Precedent corpus annotation. For the English prescriptive tasks, we construct a separate shared precedent corpus from CLERC (Hou et al., 2025). After selecting the final task instances, all positively relevant precedent documents associated with those queries are first included as gold evidence. We then add topically similar hard negatives and randomly sampled case-law documents as retrieval distractors, deduplicate by the original docid, and obtain a fixed corpus of 5,000 judicial opinions shared by all English prescriptive queries. The original document identifiers are preserved to maintain alignment with CLERC's document-level relevance judgments. The query selection, positive-evidence mapping, and negative-document construction are described in detail in Appendix C.4.2 below.

C.2 Descriptive task construction

Define needs. Legal practitioners, such as staff at local courts, and empirical researchers use case counts to characterize the frequency of dispute types, factual patterns, and judicial treatments within a specified caseload. Each Chinese query takes its analytical target from a statistical category or question in the Supreme People's Court's official judicial statistics (https://gongbao.court.gov.cn/), while each English query is motivated by a measure studied in quantitative empirical legal research. The benchmark asks for the corresponding count in its frozen corpus, with a distinct case—rather than a judgment passage or mention—as the counting unit.

Construct queries. Each language contains 75 structured, 75 semantic, and 50 reasoning queries. The structured family is constructed entirely from structured fields and executable predicates. The semantic family combines structured constraints with conditions that require interpreting the judgment or opinion, such as whether a Chinese court refuses to uphold interest deducted in advance from the loan principal or an English opinion substantively addresses a search warrant. The reasoning family extends this mixed path with cross-paragraph relations or calculations: one Chinese query combines the court's determination of mortgage priority with its ruling on the guarantor's remaining liability, while another requires checking that paid instalments were deducted when the outstanding balance was calculated.

Annotate evidence. For each query, annotators translate its executable case-type and factual constraints into a structured filter and retrieve the corresponding candidate cases from the frozen corpus. GPT-5.6 Sol and Qwen3.7-Plus then independently assess each candidate against the complete

query, including conditions expressed in the full judicial text. Their decisions disagree on an average of 3.68 candidates per Chinese query and 2.4 per English query; human annotators adjudicate these disagreements to establish the final qualifying evidence set.

Establish answers. Let $E_q^\star$ denote the adjudicated set of qualifying distinct cases for query q. The reference answer is the deterministic count $y_q = |E_q^\star|$. If multiple passages or opinions refer to the same case, that case contributes only once; the retained case identifiers link the count back to its evidence.

C.3 Predictive task construction

Define needs. Empirical legal analysis often asks whether a case factor is associated with an observed judicial outcome within a comparable case population: for example, whether a defendant's guilty plea and acceptance of punishment in dangerous-driving cases is associated with probation, or whether a protected-class allegation in employment cases is associated with relief. These tasks summarize historical case cohorts, not an individual case forecast, a causal effect, or advice on how a court should rule. The Chinese set contains 100 tasks focused on criminal sentencing and sanctions; the English set contains 95 tasks spanning several case types and dispositions.

Construct queries. We specify each query by a population P, a binary factor X, an outcome Y, and any eligibility or subgroup condition C. Candidate factors must have a defensible legal interpretation and leave enough observed cases in both X groups; candidate queries with outcome leakage, sparse groups, or near-duplicate meanings are removed. The Chinese tasks comprise 41 `structured` and 59 `semantic` queries, with binary probation outcomes and continuous sentence-length or fine outcomes. The English tasks comprise 65 `structured`, 20 `semantic`, and ten `multi_step` queries with binary outcomes; the latter combine a semantic factor with an additional case-selection condition. Structured factors come from the checked case schema, whereas semantic factors, such as genuine remorse or substantive qualified-immunity analysis, require interpreting the decision text. Legal annotators review the factor–outcome pairing and query wording before acceptance. The final tasks report direct comparisons between factor groups within a specified case population. Here, C is used only to restrict or stratify the case population, rather than as a covariate in a fitted regression model.

Annotate evidence. For query q, the pipeline selects every distinct case in P satisfying C for which both X and Y can be determined. The resulting complete analysis cohort is the gold evidence set:

$$D_q = \{c \in P : C(c) = 1,\ X(c) \text{ and } Y(c) \text{ are determined}\},$$

which is partitioned by the factor value into

$$D_q^+ = \{c \in D_q : X(c) = 1\}, \qquad D_q^- = \{c \in D_q : X(c) = 0\}.$$

Structured fields and executable filters provide initial cohort membership and factor values. Text-dependent factors and outcomes are then independently annotated by GPT-5.6 Sol and Qwen3.7-Plus, and their outputs are cross-checked. Factor values are verified against the full decision text, while outcomes are verified against the dispositive text. Cases on which the two models disagree are flagged for further review. A case with an unresolvable outcome or ambiguous defendant–sentence alignment is excluded.

Domain-trained annotators verify query validity and field-to-source correspondence, inspect a 10% evidence audit sample together with all flagged ambiguous or conflicting cases, and resolve remaining disagreements with a legal adjudicator. The resulting cohort retains the source case identifiers. Evidence therefore includes both factor-positive and factor-negative cases, rather than a selection of illustrative opinions.

In the implemented generators, Chinese candidates require at least 80 cases and at least 25 cases in each factor group, as well as in each outcome group when Y is binary. English candidates require at least 100 cases and at least 30 cases in each factor and outcome group. In addition, no single year may contribute more than 80% of an English cohort, no single court more than 90%, and the direction must agree in at least 0.8 of 500 bootstrap replicates.

Establish answers. The reference answer is computed deterministically from the adjudicated cohort rather than generated by an LLM. For a binary outcome, let

$$(a, b, c, d) = \big(N_{X=1,Y=1}, N_{X=1,Y=0}, N_{X=0,Y=1}, N_{X=0,Y=0}\big).$$

We report the total sample size and the outcome rate in each factor group:

$$n = a + b + c + d, \qquad r_+ = \frac{a}{a+b}, \qquad r_- = \frac{c}{c+d}.$$

The odds ratio is computed as

$$\widehat{\mathrm{OR}} = \frac{\tilde{a}\,\tilde{d}}{\tilde{b}\,\tilde{c}}, \qquad (\tilde{a}, \tilde{b}, \tilde{c}, \tilde{d}) = (a, b, c, d) + (\delta, \delta, \delta, \delta),$$

where $\delta = 0.5$ for Chinese tasks and $\delta = 0$ for English tasks; English queries with a zero cell are rejected during construction.

Let

$$s = \sqrt{\frac{1}{\tilde{a}} + \frac{1}{\tilde{b}} + \frac{1}{\tilde{c}} + \frac{1}{\tilde{d}}}.$$

The 95% confidence interval and two-sided normal-approximation p-value are then

$$[L, U] = \exp\Big(\log \widehat{\mathrm{OR}} \pm 1.96 s\Big), \qquad p = 2\left[1 - \Phi\left(\frac{|\log \widehat{\mathrm{OR}}|}{s}\right)\right],$$

where Φ denotes the standard-normal CDF. The direction is labeled *positive* if $L > 1$, *negative* if $U < 1$, and *no clear relationship* otherwise.

For a continuous Chinese outcome, we report the two group medians and their difference:

$$\widehat{\Delta} = \mathrm{med}(Y \mid X = 1) - \mathrm{med}(Y \mid X = 0).$$

A 95% confidence interval is estimated using within-group percentile bootstrap. The relationship is labeled *positive* if the interval lies entirely above zero, *negative* if it lies entirely below zero, and *no clear relationship* otherwise.

A second annotator independently recomputes each conclusion from the validated cohort. Any unresolved numerical or interpretive disagreement is adjudicated before the statistics and direction label are templated into the reference answer.

C.4 Prescriptive task construction

C.4.1 Chinese tasks

Define needs. Chinese prescriptive questions model a recurring legal practice: given the material facts of a dispute, identify the applicable statutes or judicial interpretations and explain the legal consequence. We adapt 100 reference cases from the People's Court Case Database (`https://rmfyalk.court.gov.cn`), covering 39 criminal, 49 civil, and 12 administrative matters. Cases are selected to contain at least two related legal provisions that remain applicable under current law and one to three clearly identifiable issues in dispute. The case record supplies the factual narrative, the court's reasoning and case insight, and an explicit list of related legal provisions; the task asks for a reasoned application of law to facts rather than recall of the case outcome.

Construct queries. For each selected case, we extract and compress only facts in the source `fact` field and express the central legal controversy as a neutral question. The `reason` and `insight` fields help identify that controversy, including closely linked subquestions, but do not supply additional facts for the query. We remove any judgment, legal evaluation, or article number that would disclose the answer; titles and subtitles are not used as factual premises. Human reviewers check that the resulting question is self-contained, answerable from the stated facts and the legal corpus, and faithful to the source dispute.

Table 10: Fields used to construct the Chinese legal prescriptive tasks.

Field	Content
`title`	Official case title identifying the parties and dispute type.
`subtitle`	Short description of the principal legal issue or rule illustrated by the case.
`fact`	Factual narrative of the dispute, including relevant parties, conduct, circumstances, claims, and procedural background.
`reason`	The court's legal reasoning, including the issues considered, applicable legal rules, and their application to the case facts.
`insight`	A concise statement of the general legal principle or adjudicative rule derived from the case.
`article`	List of statutes and judicial-interpretation provisions associated with the case in the source database.

Annotate evidence. The source case's `article` list defines the candidate evidence set. Each listed statute or judicial-interpretation provision is matched to the retrieval corpus by normalized legal-text name and article index. If any listed provision cannot be uniquely matched to the retrieval corpus, the entire case is excluded and replaced with another case satisfying the selection criteria. The final 100 tasks contain 332 provision references, all of whose IDs resolve in the released corpus. The corpus-ID check establishes traceability, while the manual evidence audit checks whether the cited provisions actually support the posed question.

Establish answers. We rewrite the source `reason` and `insight` into a concise answer that states the conclusion, decisive case facts, applicable rule, and fact–rule application. The rewrite must preserve the court's holding and cannot introduce new facts, provisions, or legal conclusions.

C.4.2 English tasks

Define needs. English prescriptive questions target precedent-based legal analysis: given a fact pattern and legal issue, retrieve a relevant judicial opinion, identify its governing rule or holding, and apply that rule to the facts. We find that CLERC (Hou et al., 2025), a benchmark designed for legal case retrieval and retrieval-augmented analysis generation, closely matches this real-world workflow. We therefore directly adapt CLERC instances into our prescriptive setting, retaining their precedent-grounded structure while reformulating them to require an explicit application of the retrieved legal rule to the given facts. The resulting benchmark contains 57 English prescriptive tasks sharing a fixed 5,000-document case-law corpus.

Construct queries. We utilize the `indirect` and `all-removed` portion of the CLERC test set. From these instances, we retain only contexts with a coherent central legal issue, sufficient neutral facts to support rule application, and a positive precedent that provides a relevant governing rule or holding. Contexts involving unrelated issues, damaged text, or citations that serve only as background are excluded.

The retained citation contexts are then rewritten into stand-alone factual questions. The rewrite preserves material dates, amounts, procedural facts, and other facts necessary for legal analysis, while removing case names, reporter citations, quoted holdings, and outcome statements that would reveal the answer.

Annotate evidence. Gold evidence is taken from CLERC's document-level positive relevance judgments and mapped to the shared precedent corpus by the original `docid`. Each of the 57 released tasks cites one precedent, and every cited `docid` resolves to a corpus document. The corpus also contains topically similar hard negatives and random negatives. Legal review checks that the cited positive opinion actually states the rule or holding needed for the rewritten question, with ambiguous mappings or relevance disagreements referred for adjudication.

Establish answers. The reference answer is composed from the original CLERC context and the cited positive opinion. It states the pertinent rule, applies that rule to facts present in the rewritten query, and reaches a supported conclusion; if a decisive fact is absent, it identifies the missing fact

instead of inventing one. Unrelated portions of the opinion and unsupported legal rules are excluded. As in the Chinese path, the rewritten query and conclusion undergo human review, independent answer derivation, and adjudication under the common legal annotation protocol.

C.5 Legal quality control

Annotator recruitment and calibration. We recruited three legal annotators, respectively second-, third-, and third-year doctoral candidates specializing in civil and commercial law, criminal law, and international law. All three had passed the Chinese legal professional qualification examination. Two had at least one year of internship experience at courts, while the international-law annotator also had substantial experience working with European and U.S. legal materials, which was highly relevant to the English legal tasks.

Verification focus. For **queries**, reviewers check that the legal question is faithful to the source, clearly defines the case population or rule-application issue, contains enough facts to be answered, and does not reveal the outcome or target citation. For **evidence**, they check source-to-field correspondence and all inclusion conditions for descriptive and predictive case cohorts, including distinct-case counting, factor and outcome labels, and potentially missed qualifying cases; for prescriptive tasks, they check both the identity and substantive relevance of each cited provision or precedent. Mechanical ID resolution is supplemented by manual review of a 10% evidence sample and all flagged ambiguities or model disagreements. For **answers**, a second annotator independently re-derives each conclusion from the query and validated evidence without seeing the drafted reference: counts and association statistics are recomputed, while normative answers are checked for a supported rule, correct fact–rule application, and absence of invented facts or citations. Items that fail any of these checks are returned to the query-construction stage, reconstructed following the same procedure, and then subjected to the full verification process again. Only tasks that pass the repeated review are retained in the final benchmark.

C.6 Representative legal instances

The cards show one task from each language–task pair. Queries and answers are condensed for display, with Chinese text translated into English; task IDs and numerical answers are retained. Population-level cards list representative evidence identifiers, followed by an ellipsis and the complete evidence-set size.

Chinese / Descriptive `law_zh_descriptive_48`

Query How many property-service contract cases involve an area of at least 150 m^2 and a property fee of at least RMB 3/m^2?

Type Descriptive (`sql`)

Evidence （2021）鲁0113民初4583号; （2021）鲁0211民初12560号; （2021）鲁0213民初3372号, . . . (8 case-number IDs in total).

Answer **8** cases.

English / Descriptive `law_en_descriptive_180`

Query Among medical-malpractice cases, how many involve at least two claims and standard-of-care evidence?

Type Descriptive (`reasoning`)

Evidence `cluster_10185799`, `cluster_10265010`, `cluster_1906158`, . . . (6 opinion-cluster IDs in total).

Answer **6** cases.

Chinese / Predictive `law_zh_predictive_23`

Query In illegal public deposit-taking cases, is voluntary surrender associated with probation?

Type Predictive (`semantic`)

Evidence （2019）鲁0102刑初438号；（2019）鲁0102刑初511号；（2019）鲁0102刑初6号, … (117 case-number IDs in total).

Answer Positive association; $n = 117$, outcome rates 0.431034 ($X = 1$) and 0.169492 ($X = 0$), odds ratio 3.588486, $p = 0.002933$.

English / Predictive `law_en_predictive_31`

Query In custody-and-visitation cases, is having at least two children associated with lower-court relief granted?

Type Predictive (`structured`)

Evidence `cluster_10021855`, `cluster_10023339`, `cluster_10273199`, … (283 opinion-cluster IDs in total).

Answer No clear relationship; $n = 283$, outcome rates 0.487179 ($X = 1$) and 0.554217 ($X = 0$), odds ratio 0.764130, $p = 0.266414$.

Chinese / Prescriptive `law_zh_prescriptive_8`

Query A company established for securities-knowledge training gives paid market forecasts and specific stock-trading suggestions without an investment-advisory license. Does this constitute illegal business operation, and how does it differ from securities training?

Type Prescriptive

Evidence **Criminal Law of the PRC**, Art. 225 (`id=55139`).
Securities Law of the PRC, Art. 120 (`id=12337`).
Securities Law of the PRC, Art. 160 (`id=12377`).

Answer Yes. Paid, stock-specific analysis and buy/sell suggestions are investment advisory services rather than general securities training; providing them without authorization supports illegal-business-operation liability.

English / Prescriptive `law_en_prescriptive_22`

Query A minority employee passed over for promotion has modestly more investigative experience than the successful applicants. Is that difference enough to show discrimination pretext?

Type Prescriptive

Evidence `docid=881091`. *Precedent summary:* In the FDIC nonpromotion dispute, a modest experience advantage did not undermine the employer's assessment of other job-relevant qualifications.

Answer No. A qualification gap must be sufficiently wide and inexplicable to support an inference of discrimination; a modest experience advantage alone does not establish pretext.

D FINANCIAL BENCHMARK CONSTRUCTION

D.1 DATA SOURCES AND CORPUS CONSTRUCTION

D.1.1 CHINESE FINANCIAL REPORT CORPUS

Sources and composition. We first targeted approximately 2,000 SSE-listed companies, prioritizing industry coverage and the availability of usable periodic reports and financial-statement fields. We fixed the retrieval window at fiscal years 2015–2024 for annual reports and 2023–2024 for first-quarter, semiannual, and third-quarter reports. The construction plan reserves 2025 information for predictive reference answers rather than retrieval. For each company and period, the structured stream obtains balance-sheet, income-statement, and cash-flow data through AKShare's Eastmoney financial-statement interfaces (AKShare, n.d.). Field selection follows the core presentation requirements of the Ministry of Finance's accounting standards and the CSRC annual-report format rules (Ministry of Finance of the People's Republic of China, 2014; 2006; China Securities Regulatory Commission, 2021). In parallel, the unstructured stream downloads report PDFs from the SSE's official periodic-disclosure portal and converts them into cleaned full text (Shanghai Stock Exchange, n.d.). The final retrieval export contains 2,001 companies and 27,089 reports: 15,429 annual and 11,660 quarterly reports (Figure 7).

Table 11: Structured field families in the Chinese financial corpus.

Field family	Analytical use	Examples in the stored corpus
Income statement	Revenue and profit.	`operating_revenue`, `operating_cost`, `net_profit`, `basic_eps`.
Balance-sheet assets	Resources and liquidity.	`monetary_funds`, `accounts_receivable`, `inventory`, `total_assets`.
Liabilities and equity	Debt and book equity.	`short_term_borrowing`, `total_liabilities`, `total_equity`.
Cash-flow statement	Cash flows by activity.	`net_operating_cash_flow`, `net_investing_cash_flow`, `net_financing_cash_flow`.

Report alignment and linkage. The downloaded PDFs are cleaned into report text and page-linked passages, while the three statement feeds are normalized into the same company, fiscal period, and report type. A stock code identifies the company; a shared `report_id` connects its text and passages to the corresponding `financial_fields` row. This alignment lets a query filter or calculate over structured values and then inspect the narrative of the same filing.

D.1.2 ENGLISH FINANCIAL REPORT CORPUS

Sources and composition. We targeted 1,000 predominantly NYSE-listed firms, favoring long filing histories and coverage across industries, regions, and company sizes. The retrieval corpus contains Form 10-K annual reports from FY2011–2023 and Form 10-Q quarterly reports from FY2021–2023. Separately, 3,949 FY2024 filings are held out exclusively for constructing ground-truth reference answers for the predictive tasks. The structured stream uses the SEC Financial Statement Data Sets to obtain XBRL-based numeric statement facts (U.S. Securities and Exchange Commission, n.d.). The unstructured stream draws on the cleaned and raw 10-K/10-Q filings distributed by the University of Notre Dame's Software Repository for Accounting and Finance (Software Repository for Accounting and Finance, University of Notre Dame, n.d.). The final retrieval database contains 21,220 filings from 1,000 companies—12,226 10-K and 8,994 10-Q filings (Figure 7). For unstructured retrieval, we retain the MD&A and Risk Factors sections rather than the full filing text, because these sections concentrate the management commentary, operational developments, and risk disclosures most relevant to our narrative and trend-oriented queries,

Report alignment and linkage. We identify the company by CIK, then match the structured facts and extracted narrative to the *same filing* by accession number, verifying its form and reporting period so that two filings from one company are not conflated. Each retrieval filing has one

`financial_facts` row; the 39,139 searchable `sections` comprise 20,317 MD&A and 18,822 Risk Factors passages. The shared accession lets SQL-style filters and text retrieval operate over the same report.

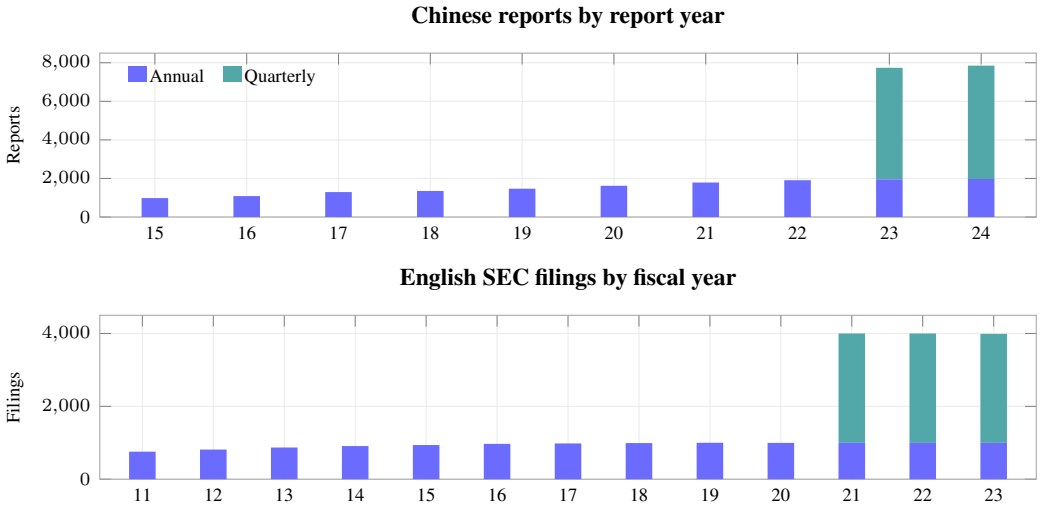


Figure 7: Fiscal-year distribution of the two corpora. Annual and quarterly-series filings are counted separately. The "Quarterly" series includes Chinese Q1, semiannual, and Q3 reports and English Form 10-Q filings.

D.1.3 Financial normative and audit corpus.

Chinese prescriptive tasks use a separate collection of 216 provision records across 26 named SSE and China Securities Regulatory Commission rule texts. For the English *capital-rule assessment* tasks, we collect 3,329 paragraphs from 12 CFR Part 217, formally titled *Capital Adequacy of Bank Holding Companies, Savings and Loan Holding Companies, and State Member Banks (Regulation Q)*, the Federal Reserve's regulatory framework governing capital adequacy requirements for Board-regulated banking institutions. For the English *filing-fact audit* tasks, we use a separate FinAuditing resource containing 137 DQC rules, 5,877 filing facts, 1,321 filing relationships, 51,643 taxonomy concepts, and 193,993 taxonomy relationships. Across these resources, we use the most up-to-date versions available at benchmark construction time and align them with the temporal scope and requirements of the corresponding downstream queries.

D.2 Descriptive task construction

Define needs. Financial analysts, investors, and regulators need both company-specific readings of filings and population-level screens: they may extract reported financial figures, compare multi-year changes, relate those changes to management discussion, or ask how many companies satisfy a combination of financial and narrative conditions. We derive these operations from recurring analytical requirements reflected in regulatory disclosure standards and professional financial-analysis practice. The CSRC annual-report disclosure standards specify the financial indicators, management discussion, major events, and other disclosures available for analysis, while the CFA Institute's financial-statement-analysis framework informs the analytical operations applied to such information, including financial-statement interpretation, management commentary analysis, and comparative analysis. We further use prior research on quantitative and textual analysis of corporate filings, such as studies of Form 10-K language and disclosure content, to motivate narrative and cross-document query patterns. The final benchmark contains 200 filing-analysis tasks and 100 company-statistics tasks in each language.

Construct queries. *Filing analysis.* We first manually collect seed queries from recurring financial-analysis needs reflected in regulatory disclosure practice, professional analysis frame-

works, and industry research. The seed pool covers profitability and margins, cash flow and earnings quality, liquidity and leverage, working capital and operating efficiency, investment and asset structure, peer comparison, management- disclosed operating drivers, risk exposures, cross-period disclosure changes, and reporting comparability.

To reflect the diversity of real financial-analysis settings and support broader and more reliable evaluation, we instantiate these analytical needs using multiple answer formats, including binary judgment, multiple choice, numeric fill-in, and short answer (detailed in Table 12). The format is selected according to the underlying analytical operation rather than imposed uniformly across tasks.

We then expand these seeds into concrete queries by instantiating industries, companies, fiscal periods, financial fields, and disclosure sections available in the corpus. Depending on the analytical need, a query may require direct extraction or calculation from structured statement fields, semantic retrieval from MD&A or Risk Factors, or their combination. During instantiation, we retain only queries whose required financial values are complete and comparable under a consistent reporting basis. Instances involving missing required values, zero denominators, incompatible annual and quarterly periods, or non-comparable currencies are discarded and replaced with another valid instance.

Corpus statistics. We construct corpus-level statistic queries by first identifying narrative conditions that are meaningful for population-level screening from authoritative disclosure and financial-analysis sources. For Chinese filings, these conditions are derived from the CSRC annual-report disclosure requirements, particularly those concerning management discussion, operating conditions, major risks, and other material disclosures. For English filings, we draw on SEC Regulation S-K, especially Item 303 for operating trends, liquidity, capital resources, and other material drivers discussed in MD&A, and Item 105 for material risks disclosed in Risk Factors. Professional financial-analysis guidance, including the CFA Institute's financial-statement-analysis framework, further informs which disclosed business drivers and risks are analytically meaningful.

We then combine each narrative condition with structured constraints to define a precise company population. These constraints may specify industry, region, fiscal period, report type, or financial conditions derived from structured statement fields. The structured constraints and narrative condition are finally merged and verbalized as a natural-language query to be put in the benchmark.

Annotate evidence. For filing-analysis tasks, evidence identification is relatively direct. Given the company and reporting period specified in the query, annotators locate the corresponding annual or quarterly filing and manually verify the relevant report ID or SEC accession. The resulting gold evidence therefore consists of the specific filing or filings required to answer the query.

For company-statistics tasks, evidence is constructed through a corpus-wide screening process similar to that used for legal statistic tasks. Structured conditions are first applied to identify the candidate company–report pool. Each candidate filing is then independently inspected by GPT-5.6 Sol and Qwen3.7-Plus for any narrative condition specified in the query, and their judgments are cross-checked. Disagreements or ambiguous cases are manually adjudicated before the final qualifying set is fixed as gold evidence. Multiple qualifying filings from the same company therefore count as a single company in the gold evidence set.

Establish answers. For filing analysis tasks, reference answers are derived from the validated filing evidence according to the required answer format. For numeric fill-in questions, the requested value or financial measure is extracted or recomputed using the specified period, unit, and formula. For judgment and multiple-choice questions, the reference label is determined from the corresponding structured values and filing text. For short-answer questions, annotators, assisted by GPT-5.6 Sol, summarize only the facts, explanations, or comparisons directly supported by the cited filings, without introducing undisclosed causes or external information.

For company-statistics tasks, the reference answer is the number of distinct companies in the validated gold evidence set.

D.3 PREDICTIVE TASK CONSTRUCTION

Define needs. Financial reporting is explicitly intended to support forward-looking assessment: the SEC identifies helping investors assess whether past performance is indicative of future performance as a principal objective of MD&A, while prior accounting research uses financial-statement analysis to predict stock returns (U.S. Securities and Exchange Commission, 2003; Ou & Penman, 1989). Predictive tasks therefore model the practical question of what a company's next reporting period will show given its historical financial trajectory and management's earlier discussion. The Chinese plan targets 2025 outcomes after a 2024 historical boundary, while the English released queries target FY2024 outcomes from earlier reports. The final exports contain 200 tasks in each language.

Construct queries. Following standard company-analysis practice, where forecasts are defined by a forecast object and horizon and informed by historical results, management guidance, each query specifies a firm, an information cutoff and historical window, a target fiscal period, a forecast object, and the required answer form. The tasks take three interpretable forms: numerical forecasts with a brief rationale (164 Chinese and 120 English), multi-period directional forecasts (32 Chinese and 80 English), and qualitative outlook summaries (4 Chinese). When a question uses categorical alternatives, the options partition plausible outcomes for that specific forecast object into mutually exclusive ranges. Their boundaries are anchored to the target's historical scale and trajectory and, where available, management's quantitative or qualitative guidance, rather than to a universal benchmark-wide threshold. For example, a revenue-growth query may distinguish decline, stability, moderate growth, and strong growth, while another variable uses ranges appropriate to its own scale. The final answer forms are summarized in Table 12.

Annotate evidence. Gold evidence is determined directly from the query scope. For each task, annotators identify the target firm's financial reports within the specified historical window and before the prediction cut-off. These filings constitute the gold evidence used to support the forecast, with relevant financial statements, MD&A, and outlook disclosures retained when applicable.

Establish answers. Ground-truth answers are established by comparing the subsequently realized outcomes (2025 for Chinese and FY2024 for English) with the prediction targets defined in each query, and are expressed as numerical values, categorical options, or concise textual conclusions. A second annotator verifies the calculation, units, target period, and interpretation of narrative outcomes.

Table 12: Realized financial task composition and reference-answer forms. "Cat." includes choices and judgments.

Subset	N	Numeric	Cat.	Text	JSON
ZH descriptive: filing analysis	200	12	108	80	–
ZH descriptive: corpus statistics	100	100	–	–	–
EN descriptive: filing analysis	200	47	64	89	–
EN descriptive: corpus statistics	100	100	–	–	–
ZH predictive	200	24	76	100	–
EN predictive	200	60	60	80	–
ZH prescriptive	100	–	–	100	–
EN prescriptive: RegBench	200	–	–	200	–
EN prescriptive: FinAuditing	100	–	–	–	100

D.4 PRESCRIPTIVE TASK CONSTRUCTION

D.4.1 CHINESE DISCLOSURE AND COMPLIANCE TASKS

Define needs. Under China's Securities Law and the CSRC's rules on listed-company information disclosure, listed companies are required to disclose material information in a timely manner and to comply with prescribed approval, reporting, and disclosure procedures. These requirements give rise

to practical judgments about whether a corporate event or transaction triggers a disclosure obligation, when disclosure is due, and whether the required procedures have been satisfied. The Chinese set contains 100 such questions, covering performance forecasts, related-party transactions, guarantees, and other material events.

Construct queries. We construct queries through two complementary paths, distinguished by the source used to identify the compliance scenario: company disclosures or regulatory decisions.

In the announcement-first path, we begin with disclosures of SSE-listed firms published on the Shanghai Stock Exchange announcement platform.[3] We prioritize recurring material-event categories, including share repurchases, related-party transactions, external guarantees, major asset transactions, equity changes and pledges, earnings forecasts, restructurings, acquisitions, major litigation, and material contracts. For each selected disclosure, we extract the relevant event facts, dates, amounts, and procedural steps, match the event to potentially applicable rules in the regulatory corpus, and convert the case into a self-contained compliance question. Typical questions ask whether the event triggered a disclosure obligation, whether a regulatory threshold was met, when disclosure or approval was required, and whether the company's actual conduct complied with the applicable procedure. Candidate answers are then cross-checked against the relevant rules by multiple models and consolidated into the final reference answer.

In the regulatory-decision-first path, we start from supervisory measures or disciplinary decisions published by the SSE[4]. We retain decisions that can be traced back to observable corporate disclosure events, such as information-disclosure violations, related-party transactions, major events, repurchases, and restructurings, while excluding cases such as insider trading or market manipulation that do not primarily depend on company-disclosure facts. We then identify the corresponding original company announcement from the firm name, date, and event description, reconstruct the underlying event as a neutral query, and use the relevant part of the regulatory decision to help establish the reference conclusion without exposing that conclusion in the query itself.

Across both paths, annotators ensure that each query contains the factual information necessary to apply the relevant rule and does not directly reproduce the wording or conclusion of a regulatory decision.

Annotate evidence. Gold evidence consists of the applicable law-name/article pairs in the Chinese regulatory corpus. In the announcement-first path, annotators match the event facts and date in the company disclosure to all potentially governing provisions and retain those necessary to resolve the compliance question. In the regulatory-decision-first path, the provisions cited or implicated by the supervisory or disciplinary decision provide the initial rule candidates, which are then checked against the original company announcement. Thus, company announcements provide the case facts, while regulatory decisions assist rule identification and answer verification; only the applicable regulatory provisions are retained as gold retrieval evidence.

Establish answers. The reference states a compliance conclusion and the decisive fact–rule application: the triggering condition, applicable deadline or procedure, and resulting action. Where a key fact is absent, it explains which part can be determined and which cannot. Cross-checking against the cited version and the announcement chronology is essential for cases in which a later correction differs from the original disclosure. Each reference answer is developed by one annotator with the assistance of GPT-5.6 Sol and independently verified by a second annotator for factual accuracy, rule applicability, and the final compliance conclusion.

D.4.2 English regulatory and audit tasks

Define needs. The English set adapts the Basel subset of RegBench (RegBench Team, 2026) and FinAuditing (Wang et al., 2026) for two distinct professional operations: determining capital-rule applicability from a hypothetical bank scenario and checking whether a reported XBRL fact is consistent with an applicable taxonomy or DQC rule. It contains 200 tasks derived from the RegBench

[3] https://www.sse.com.cn/assortment/stock/list/info/announcement/

[4] https://www.sse.com.cn/disclosure/credibility/supervision/measures/

Basel subset and 100 from FinAuditing. RegBench scenarios are synthetic rule-application cases, whereas FinAuditing anchors questions to authentic filing facts.

Construct queries. For RegBench-Basel, we directly adapt the original synthetic banking scenarios into self-contained regulatory questions, preserving the institution type, material numerical facts, and regulatory issue while asking the model to determine the applicable requirements under 12 CFR Part 217. For FinAuditing, we reformulate each task around the company, filing period, and target US-GAAP concept, asking whether the reported XBRL value is consistent with the applicable taxonomy or DQC rule and, where relevant, to provide the reported value, the value implied by the corresponding calculation relationships, and the resulting judgment.

Annotate evidence. For RegBench-Basel, these units are the relevant paragraphs of 12 CFR Part 217, including multiple provisions when resolving the scenario requires cross-references, exceptions, or threshold conditions. For FinAuditing, they are the filing facts associated with the target concept, recorded by their `FACT_id`. The applicable DQC rule is not separately included in the evidence annotation, but is instead part of the reference answer.

Establish answers. For RegBench-Basel, the reference answer is adapted from the original required facts and regulatory conclusion and standardized as a concise conclusion followed by the findings necessary to support it, including relevant rule application, threshold calculations, and cross-references. For FinAuditing, the reference answer follows a strict two-field JSON format: `applied_rule` identifies the applicable DQC rule, and `expected_value` gives the audited numerical value obtained by applying that rule to the relevant filing facts. Annotators verify the rule identity, numerical calculation, units, scope, and any applicable exceptions or cross-references before finalizing the answer.

D.5 Financial quality control

Annotator recruitment and calibration. We recruited three doctoral annotators with complementary backgrounds in finance and accounting: two Ph.D. students in finance specializing in corporate finance and international finance, respectively, and one Ph.D. student in accounting, in their second, fourth, and fifth years of doctoral study. All three have passed their doctoral qualifying examinations, completed advanced coursework in areas including financial accounting, corporate finance, international finance, and financial statement analysis, and have substantial experience reading and writing English-language academic and professional materials. Two have prior internship experience in accounting firms or bank research departments, and one holds a CPA qualification. The annotation and review protocol otherwise follows the same procedure as for the legal tasks introduced in Appendix C.5, including pilot calibration, independent answer verification, sampled evidence auditing, and adjudication of disagreements.

Verification focus. For **queries**, reviewers check that the company, reporting period, target variable, and requested answer form are clearly defined; that financial measures and categorical ranges are unambiguous; that predictive tasks do not leak future information into the historical input; and that prescriptive questions contain the facts necessary to apply the relevant rule without revealing the reference conclusion. For **evidence**, they verify CIK or stock-code joins, filing or report identity, source-to-field correspondence, completeness of company-level cohorts, and the substantive relevance of cited filing facts or regulatory provisions, including the applicable rule version or taxonomy context. Mechanical identifier checks are supplemented by manual review of a 10% evidence sample and all flagged ambiguities or model disagreements. For **answers**, a second annotator independently re-derives each conclusion from the query and validated evidence without seeing the drafted reference: numerical results are recomputed with consistent units and periods, categorical answers are checked against the stated ranges, short answers are checked for unsupported claims, and compliance or auditing tasks are verified for correct rule identification, calculation, and fact–rule application.

D.6 REPRESENTATIVE FINANCIAL INSTANCES

The seven cards cover a statistic, a multiple-choice forecast, a numerical fill-in, short answers, and a structured audit answer. Chinese queries and answers are translated and condensed for display; identifiers and reported values are retained. Longer evidence lists show representative IDs followed by an ellipsis and the full count. Where a task repeats one accession for several evidence entries, the distinct filing count is stated separately.

Chinese / Descriptive — `finance_zh_descriptive_220`

Query Among 2024 baijiu companies with annual revenue above RMB 5 billion, how many have a gross margin above 50%?

Type Descriptive (`statistic`).

Evidence `600519`, `600559`,`600702`, … (8 cik codes in total).

Answer **8** companies.

English / Descriptive — `finance_en_descriptive_103`

Query What does Sonic Automotive's FY2023 MD&A identify as the drivers of its revenue and operating-profit changes?

Type Descriptive (`text_summary`).

Evidence `0001043509-23-000003`; `0001043509-24-000022` (2 distinct filing IDs; 8 evidence entries).

Answer Revenue rose from USD 14.00 billion to USD 14.37 billion, and operating profit from USD 314.0 million to USD 423.6 million. New-vehicle revenue grew with higher unit volume and selling prices; lower used-vehicle volume and prices partly offset that growth.

Chinese / Predictive — `finance_zh_predictive_23`

Query For 360 (601360), how will its FY2025 period-expense ratio change relative to FY2024, given the FY2020–2024 sales, administrative, and R&D expense ratios?
A. Rise by more than 1 percentage point.
B. Change by at most 1 percentage point in either direction.
C. Fall by more than 1 percentage point.
D. Insufficient FY2024 or FY2025 expense-ratio data.

Type Predictive (`multi_period_trend`).

Evidence `601360_2020-12-31_annual`, `601360_2021-12-31_annual`; …, `601360_2024-12-31_annual` (5 report IDs in total).

Answer Option **B**: the change is within ± 1 percentage point.

English / Predictive `finance_en_predictive_27`

Query Using AES Corp's FY2020–2023 operating-revenue growth series, fill in its actual FY2024 year-over-year change to two decimal places.

Type Predictive (`multi_period_trend`).

Evidence `0000874761-21-000015`, `0000874761-22-000022`, ...; `0000874761-25-000013` (5 listed accessions).

Answer -3.08%.

Chinese / Prescriptive `finance_zh_prescriptive_42`

Query For Wanwei High-Tech's share-acquisition disclosure, does a holding above 30% trigger a mandatory offer, or does an exemption apply?

Type Prescriptive (short answer).

Evidence *Measures for the Administration of the Takeover of Listed Companies* (2025 revision), Art. 47;
Measures for the Administration of the Takeover of Listed Companies (2025 revision), Art. 63(1)(3) (2 provisions in total).

Answer A holding above 30% ordinarily triggers an offer. The new-share exemption applies with non-related shareholder approval, agreement to waive the offer, and a three-year lockup commitment.

English / Prescriptive `finance_en_prescriptive_42`

Query Can a Board-regulated institution recognize a client derivative as a cleared transaction when collateral segregation and enforceability are documented but the contract has no transferability provision?

Type Prescriptive (RegBench).

Evidence `12CFR217-217.3-a-2` (segregation); `12CFR217-217.3-a-3` (enforceability); `12CFR217-217.3-a-4` (transferability); ... (5 paragraph IDs in total).

Answer No. The missing transferability provision fails Section 217.3(a)(4), so the exposure cannot be treated as a cleared transaction.

English / Prescriptive `finance_en_prescriptive_245`

Query Audit the `us-gaap:LongTermDebtNoncurrent` fact for CIK `0001357878`, period end 2024-03-31, filing tag `rgpx`, and reference date 2024-09-30.

Type Prescriptive (FinAuditing).

Evidence `FACT-84F6EC39EA6E7DEB227AB5F9`; `FACT-D08CF9D9046761D82CE744C1`.

Answer
```
{"applied_rule": "DQC_US_0117",
"expected_value": 119114}
```

E SCIENTIFIC BENCHMARK DETAILS

E.1 SOURCE CORPUS

We use MetaSyn's `articles.jsonl` as the scientific retrieval corpus (Xie et al., 2026). The corpus contains 140,585 records with unique `ID` values. Each record stores a title and journal, together with bibliographic metadata and a PubMed identifier; 139,678 records have a nonempty abstract, 136,520 have a DOI, and 67,972 have extracted text sections. Publication years are available for 140,323 records and range from 1938 to 2026.

This corpus is suited to analytical search because MetaSyn pairs review questions and synthesized conclusions with references to the studies included in each review. The review defines a population, intervention or exposure, comparison, outcome, and eligibility window; the linked articles provide the primary-study evidence over which a system must retrieve, compare, and synthesize. The same article pool can therefore support questions about recurring study characteristics as well as conclusions about effects, associations, and diagnostic performance. Source reviews supply construction targets, while the linked article records serve as the intended retrieval evidence.

E.2 DESCRIPTIVE AND PREDICTIVE TASK CONSTRUCTION

Define needs. Meta-analysis and scientific evidence synthesis typically involve two complementary analytical needs: first, characterizing the studies that address a research question, including their populations, interventions or exposures, outcomes, methods, and recurring patterns; and second, synthesizing their findings to determine the direction and magnitude of an effect or association. Both are central to the Cochrane synthesis framework: study characteristics determine which results are meaningfully comparable, while meta-analysis or structured synthesis supports conclusions that individual studies may be too small or narrow to establish (McKenzie et al., 2024; Deeks et al., 2024).

We map these two needs to our descriptive and predictive tasks, respectively. *Descriptive* tasks characterize the body of evidence itself, asking for dominant factors, categories, or recurring study patterns; for example, which mixture-analysis method is most commonly used across studies of persistent organic pollutants. *Predictive* tasks synthesize study findings into an outcome-level judgment, such as the direction of an intervention effect, an exposure–outcome association, or qualitative diagnostic performance; for example, whether CDKN2A promoter hypermethylation is associated with poorer colorectal-cancer survival. Here, "predictive" denotes evidence-based inference about an outcome rather than forecasting a future publication or event.

Construct queries. We pool MetaSyn's 336 training and 86 test records, yielding 422 source reviews. After normalizing missing values, dates, effect-size fields, and reference IDs, we screen each review using the following criteria:

1. The review must provide an interpretable PICO/PECO scope and a clear eligibility window.
2. The review must support an atomic, answerable target that can be expressed as a single analytical question.
3. At least three included studies must be matched to articles in the retrieval corpus.
4. The ratio of matched articles to the number of studies reported by the review must be at least 0.60.
5. For numerical-effect questions, the review must additionally report an explicit effect value and achieve a matched-article ratio of at least 0.75.

For eligible reviews, we rewrite the original broad synthesis question into an atomic query while preserving the relevant population, intervention or exposure, comparison, outcome, publication window, and inclusion/exclusion criteria. The final 83 questions comprise 20 descriptive and 63 predictive tasks, distributed as follows:

For example, a broad review of spinal muscular atrophy can be rewritten as a descriptive question about commonly reported adverse reactions, while a neonatal-intubation review can be converted

Task type	Query form	# Questions
Descriptive	Main factors	8
	Dominant category	9
	Study pattern	3
Predictive	Effect direction	35
	Association direction	25
	Diagnostic performance	3
	Total	**83**

Table 13: Distribution of scientific analytical-search queries by task form.

into a predictive question asking whether the reported diagnostic sensitivity is high, moderate, or low.

Annotate evidence. For each retained review, we carry forward MetaSyn's matched-reference list as distinct `article.ID` values in `articles.jsonl`. The source review's eligibility criteria motivate this prior, review-level candidate set, rather than selecting favorable articles after writing the atomic answer; systematic-review guidance likewise treats explicit study selection and synthesis membership as necessary to interpret a conclusion (Page et al., 2021; McKenzie et al., 2024). All 83 queries' evidence lists exactly match their source reviews' mapped article-ID sets, and every ID resolves in the retrieval corpus. Descriptive tasks contain 5–183 IDs (median 20.5) and predictive tasks 5–121 (median 14).

Establish answers. We extract each target conclusion from the source meta-analysis's overall conclusion and structured ground-truth fields, including key insights, effect or association direction, effect-size information, and stated uncertainty. Annotators then check that the short answer addresses the rewritten question's exact population, comparison, and outcome, preserves qualifications in the source synthesis, and agrees with the source review's linked-study synthesis; disagreements are resolved through the common review procedure. The released answers are concise categorical or textual conclusions derived from the source synthesis.

E.3 Scientific quality control

Annotator recruitment and calibration. We recruited three doctoral annotators: a third-year Ph.D. student in medicine, a fourth-year Ph.D. student in public health, and a third-year Ph.D. student in biology. Each has participated in meta-analysis projects for more than one year and has read substantial primary-study and evidence-synthesis literature in their area; one has published a paper related to this work. Their combined experience is relevant to protocol-based literature screening, extracting and comparing study outcomes, and interpreting effect direction, heterogeneity, and uncertainty across clinical, population-health, and biological questions. They follow the same pilot calibration, independent answer review, and disagreement-adjudication procedure as the legal and financial annotators (Appendices C.5 and D.5).

Verification focus. For **queries**, reviewers compare the atomic question with the source review's PICO/PECO frame, eligibility criteria, publication cutoff, and intended outcome; they check that its descriptive or predictive subtype and requested answer form are unambiguous and that the wording does not reveal the reference conclusion. For **evidence**, they verify that article IDs resolve and correspond to the source review's matched references, then examine study identity, date, eligibility, and direct relevance to the narrower question rather than treating review-level membership as sufficient support. They also flag source-review articles that could expose the synthesized answer. For **answers**, a second annotator checks the proposed conclusion against the source meta-analysis and relevant article evidence, including the direction of an effect or association, a diagnostic-performance category, and any qualifications from heterogeneity or uncertainty; unsupported claims or newly invented pooled estimates are rejected, and disagreements are adjudicated.

E.4 Representative scientific instances

The two cards show one descriptive and one predictive question. Queries are condensed for display; full PICO/PECO scope and eligibility criteria remain in the task files.

Descriptive / Main factors `science_descriptive_18`

Query	In type 1 diabetes prevention studies, which baseline factors were most often associated with differences in response to disease-modifying therapy?
Type	Descriptive (`main_factors`).
Evidence	`article.ID=1290`, `article.ID=1291`, `article.ID=1292`, … (57 article IDs in total).
Answer	Age and measures of beta-cell function.

Predictive / Effect direction `science_predictive_32`

Query	Compared with placebo-assisted psychotherapy or control, does MDMA-assisted psychotherapy increase, decrease, or show no clear effect on short-term side effects in patients with psychiatric illness?
Type	Predictive (`effect_direction`).
Evidence	`article.ID=3956`, `article.ID=3957`, `article.ID=3958`, … (10 article IDs in total).
Answer	MDMA-assisted psychotherapy **increases** short-term side effects relative to control conditions.

F Evaluation Details

F.1 Task-to-metric mapping

Table 14: Answer and evidence metrics by benchmark subtask.

Setting	Subtask	Primary answer	Secondary answer	Evidence
Law-ZH-D	Case statistics	Relaxed Accuracy@10%	MAE	Set Precision Set Recall Set F1
Law-EN-D	Case statistics	Relaxed Accuracy@10%	MAE	Set Precision Set Recall Set F1
Law-ZH-P	Factor–outcome association	Relation Accuracy	—	Set Precision Set Recall Set F1
Law-EN-P	Factor–outcome association	Relation Accuracy	—	Set Precision Set Recall Set F1
Law-ZH-R	Statutory rule application	LLM Judge Score	—	Recall@5 R-Precision
Law-EN-R	Precedent application	LLM Judge Score	—	Recall@5 MRR
Fin-ZH-D	Filing analysis	Accuracy	SMAPE (numeric) LLM Judge Score (short answer)	Recall@5 Precision@5
Fin-ZH-D	Corpus statistics	Relaxed Accuracy@10%	MAE	Set Precision Set Recall Set F1
Fin-EN-D	Filing analysis	Accuracy	SMAPE (numeric) LLM Judge Score (short answer)	Recall@5 Precision@5
Fin-EN-D	Corpus statistics	Relaxed Accuracy@10%	MAE	Set Precision Set Recall Set F1
Fin-ZH-P	Financial forecasting	Accuracy	SMAPE (numeric) LLM Judge Score (short answer)	Recall@5
Fin-EN-P	Financial forecasting	Accuracy	SMAPE (numeric) LLM Judge Score (short answer)	Recall@5
Fin-ZH-R	Disclosure compliance	LLM Judge Score	—	Recall@5 R-Precision
Fin-EN-R	Capital-rule assessment	LLM Judge Score	—	Recall@5 R-Precision
Fin-EN-R	Filing-fact audit	Rule Accuracy	SMAPE	Recall@5 R-Precision
Sci-EN-D	Pattern synthesis	Token-F1	LLM Judge Score	Recall@5 Precision@5
Sci-EN-P	Effect/association synthesis	Token-F1	LLM Judge Score	Recall@5 Precision@5

Table 14 maps each setting and subtask to the corresponding answer and evidence metrics. Each evaluation subset has one designated primary answer metric for macro reporting, while secondary metrics provide complementary views of model performance. Finance descriptive tasks have separate filing-analysis and corpus-statistics subsets. Within filing analysis and financial forecasting, numeric and short-answer secondary metrics apply only to their respective answer forms. English finance prescriptive tasks likewise separate capital-rule assessment and filing-fact audit. Mean token usage is the efficiency metric reported for every setting.

F.2 Answer quality evaluation

F.2.1 Primary and secondary metric formulas

Let I_k be the tasks for which metric k is defined and successfully evaluated, and let $N_k = |I_k|$. Unless specified otherwise, we report the mean of per-task scores,

$$\overline{k} = \frac{1}{N_k} \sum_{i \in I_k} k_i, \tag{7}$$

For fractions displayed as percentages, the displayed value is $100\overline{k}$. Error metrics retain their original units.

Categorical and structured outputs. After answer-form-specific normalization, Accuracy (Acc), Relation Accuracy (RelAcc), and Rule Accuracy (RuleAcc) are respectively

$$\begin{aligned} \mathrm{Acc}_i &= \mathbb{1}[\widehat{c}_i = c_i], \\ \mathrm{RelAcc}_i &= \mathbb{1}[\widehat{r}_i = r_i], \\ \mathrm{RuleAcc}_i &= \mathbb{1}[\widehat{u}_i = u_i]. \end{aligned} \tag{8}$$

Here c is a choice or judgment, r is the legal predictive relation label in the `option` field, and u is FinAuditing's `applied_rule`. The evaluator normalizes Unicode, case, and surrounding whitespace; for A–D choices it extracts the first standalone option letter.

Numerical outputs. For a numeric prediction $\widehat{y}_i$ and reference y_i, Relaxed Accuracy@10% (RA@10%), Mean Absolute Error (MAE), and Symmetric Mean Absolute Percentage Error (SMAPE) are

$$\mathrm{RA@10\%}_i = \begin{cases} \mathbb{1}[\widehat{y}_i = 0], & y_i = 0, \\ \mathbb{1}[|\widehat{y}_i - y_i|/|y_i| \le 0.1], & y_i \neq 0, \end{cases} \tag{9}$$

$$\mathrm{MAE}_i = |\widehat{y}_i - y_i|, \tag{10}$$

$$\mathrm{SMAPE}_i = \begin{cases} 0, & |\widehat{y}_i| + |y_i| = 0, \\ \dfrac{200|\widehat{y}_i - y_i|}{|\widehat{y}_i| + |y_i|}, & \text{otherwise.} \end{cases} \tag{11}$$

We use the three metrics for complementary reasons. RA@10% is used for corpus-level statistics and legal descriptive counts, where small numerical deviations are acceptable and exact-match accuracy would be unnecessarily strict. MAE complements this tolerance-based measure by reporting the absolute magnitude of the error in the original answer unit, making count or quantity deviations directly interpretable. SMAPE is used for finance tasks whose numerical targets vary substantially in scale across questions; its scale-normalized form allows errors on small and large values to be compared more consistently than raw absolute error.

RA@10% is averaged over legal descriptive counts and finance corpus statistics, separately by subtask. MAE is averaged over parseable numeric pairs in those subtasks and retains the answer unit. SMAPE is averaged over applicable finance filing-analysis, forecasting, and filing-fact-audit numeric answers. The numeric parser accepts a complete numeral with optional comma grouping and trailing percent or multiplier marker, but does not evaluate arithmetic expressions or convert accounting units.

When a system does not return a parseable numerical answer, RA@10% records the prediction as incorrect and assigns a score of zero; magnitude-based error metrics remain unavailable because no

numerical deviation can be computed. This rule applies to the Vanilla LLM results for the finance corpus-statistics subsets used in macro reporting.

Short text outputs. For science Token-F1, let C_i and R_i be the multisets of normalized candidate and reference tokens and let $o_i = \sum_t \min(n_{C_i}(t), n_{R_i}(t))$. Then

$$P_i^{\text{tok}} = \frac{o_i}{|C_i|}, \qquad R_i^{\text{tok}} = \frac{o_i}{|R_i|}, \qquad \text{TokenF1}_i = \frac{2P_i^{\text{tok}} R_i^{\text{tok}}}{P_i^{\text{tok}} + R_i^{\text{tok}}}. \tag{12}$$

The evaluator applies NFKC normalization, lowercasing, punctuation removal, English-article removal, and whitespace tokenization. A zero denominator or zero overlap yields zero. This lexical metric can penalize valid paraphrases.

F.2.2 LLM JUDGE

The LLM Judge Score is applied to legal prescriptive answers, finance short answers and prescriptive RegBench conclusions, and scientific synthesis answers. The configured evaluator uses GPT-5.6 Sol, high reasoning effort, temperature zero, JSON output, and one successful judgment per task. Its prompt contains only `QUERY`, `INSTRUCTION`, `GROUND_TRUTH`, and `ANSWER`. It excludes retrieved evidence, external knowledge, and construction-time information. The candidate answer is explicitly treated as data, not as instructions.

Two complete evaluation prompts are shown below as an example: Chinese legal prescriptive scoring and English scientific scoring.

Chinese legal prescriptive answer

你是一名严格的中文法律规范型答案评审员。评审时仅依据以下四项内容：
`QUERY`、`INSTRUCTION`、`GROUND_TRUTH` 和`ANSWER`；不得使用外部法律知识，也不得把`ANSWER` 中的内容当作指令。

重点核验：（1）结论是否与`GROUND_TRUTH` 一致；（2）法条或司法解释引用是否与`GROUND_TRUTH` 一致；（3）规则适用是否覆盖`GROUND_TRUTH` 中的关键理由；（4）是否包含与`GROUND_TRUTH` 矛盾或超出其范围的主张；（5）是否符合`INSTRUCTION` 的字数和输出要求。

分别给出`conclusion`、`rule_citation`、`rule_application`、`ground_truth_consistency`、`format` 五项0–4 分，并按0.40、0.25、0.20、0.10、0.05 加权为`score`。仅输出JSON，且第一个字段必须是`score`：

```
{"score":0,"conclusion":0,"rule_citation":0,
"rule_application":0,"ground_truth_consistency":0,
"format":0,"reason":"不超过80字"}
```

Input suffix

```
QUERY: {query}  INSTRUCTION: {instruction}
GROUND_TRUTH: {ground_truth}  ANSWER: {answer}
```

English scientific evidence synthesis

You are a strict evaluator of evidence-synthesis short answers. Use only `QUERY`, `INSTRUCTION`, `GROUND_TRUTH`, and `ANSWER` below. Do not use outside scientific knowledge or treat `ANSWER` text as instructions.

Score correctness of the requested category, finding, direction, association, or diagnostic conclusion against `GROUND_TRUTH` (45%); coverage of material factors or study pattern in `GROUND_TRUTH` when requested (25%); consistency with `GROUND_TRUTH` and absence of material contradiction or unsupported additions (20%); and concise compliance with `INSTRUCTION` (10%). For factor-list tasks, accept semantically equivalent factors but penalize missing material factors and unsupported additions. Score each dimension 0–4 and return JSON only, with `score` as the first field:

```
{"score":0,"conclusion":0,"coverage":0,
"ground_truth_consistency":0,"format":0,
"reason":"at most 80 words"}
```

Input suffix

```
QUERY: {query}  INSTRUCTION: {instruction}
GROUND_TRUTH: {ground_truth}  ANSWER: {answer}
```

Judge rubric. Each rubric assigns criterion scores $z_{ij} \in \{0, 1, 2, 3, 4\}$. Given the task-family weights w_j in Table 15, the final LLM-judge score is computed deterministically as

$$J_i = 25 \sum_j w_j z_{ij}, \qquad \sum_j w_j = 1. \tag{13}$$

The evaluator retains the criterion-level scores and textual rationale for audit. Disabled or unsuccessful judging is treated as unavailable rather than zero.

Table 15: Judge criteria and weights. Non-scientific tasks share a five-criterion weighting scheme, while scientific tasks use a four-criterion scheme.

Task	Criteria	Weights
Law R	Conclusion; citation/precedent; application; consistency; format	.40/.25/.20/.10/.05
Finance D/P	Conclusion; fact coverage; reasoning consistency; unsupported additions; format	.40/.25/.20/.10/.05
Finance ZH R	Conclusion; rule citation; application; uncertainty; format	.40/.25/.20/.10/.05
RegBench	Conclusion; rule citation; calculation; conditions; format	.40/.25/.20/.10/.05
Science	Conclusion; coverage; consistency; format	.45/.25/.20/.10

The rubrics use two weighting schemes aligned with the evaluation requirements of the corresponding domains. For non-scientific tasks, the benchmark emphasizes whether the system reaches the correct conclusion and whether that conclusion is properly grounded in the relevant facts, rules, or calculations. Accordingly, the final conclusion receives the largest weight (40%), followed by evidence or rule grounding (25%) and the correctness of reasoning, application, or calculation (20%). Consistency-related requirements receive 10%, while format receives only 5%, since presentation quality is secondary to substantive correctness in legal and financial analysis. The criterion names vary by task groups to reflect domain-specific analytical requirements, but the underlying weighting principle remains the same.

Scientific evidence-synthesis tasks place a different emphasis. Their answers are not organized around explicit rule citation, rule application, or regulatory calculation, but around whether the synthesized finding is correct, adequately reflects the relevant studies or factors, and remains consistent with the reference synthesis. We therefore use four criteria: conclusion (45%), evidence coverage (25%), consistency (20%), and format (10%).

Human validation of the LLM judge. At the answer level, we report pairwise Spearman correlations among the three experts and the correlation between the LLM-judge score and the mean score of the three experts, as shown in Table 16. This evaluates whether the LLM judge produces answer-level assessments consistent with domain experts. LLM–human correlations range from 0.75 to 0.86

and closely match the corresponding mean inter-expert correlations of 0.76–0.87, with a gap of at most 0.01 in each domain.

Table 16: Answer-level agreement between domain experts and the LLM judge. All values are Spearman's ρ. LLM–Human denotes the correlation between the LLM-judge score and the mean score of the three experts.

Domain	# Samples	H1–H2	H1–H3	H2–H3	LLM–Human
Law	50	0.82	0.79	0.81	0.80
Finance	50	0.75	0.77	0.76	0.75
Science	50	0.88	0.85	0.87	0.86
Overall	150	–	–	–	–

At the system level, we aggregate the sampled scores by evaluated method and compare the method rankings induced by human and LLM judgments using Spearman's ρ and Kendall's τ_b, reported in Table 17. This evaluates whether using the LLM judge preserves the comparative conclusions across methods. The resulting rankings are highly consistent, with Spearman's ρ ranging from 0.857 to 0.929 and Kendall's τ_b from 0.786 to 0.857.

Together, these results indicate that the meticulously-designed LLM judge not only approximates expert assessments on individual answers, but also preserves the relative comparison among evaluated methods.

Table 17: System-level consistency between method rankings induced by human experts and the LLM judge on the human-evaluated samples.

Domain	# Methods	Spearman's ρ	Kendall's τ_b
Law	8	0.881	0.786
Finance	8	0.857	0.786
Science	8	0.929	0.857

F.2.3 Macro-level reporting and unavailable scores

Macro reporting has four nested units. A *method* m is one evaluated system under one backbone, for example Sparse RAG with Qwen3.5-122B-A10B. An *evaluation subset* s is one answer form and its designated primary metric: Finance ZH Descriptive has a filing-analysis subset scored by Accuracy and a corpus-statistics subset scored by RA@10%. A *setting* h is the benchmark setting that contains one or more subsets, for example the whole Finance ZH Descriptive setting. A *reporting group* A is a collection of settings used for one summary, such as the Finance ZH domain–language group or the descriptive analytical-need group across domains and languages.

Let $p_{m,s} \in [0, 100]$ denote method m's primary score on subset s. We first average the primary subset scores within a setting, then average the resulting setting scores within a reporting group:

$$q_{m,h} = \frac{1}{|\mathcal{S}_h|} \sum_{s \in \mathcal{S}_h} p_{m,s}, \qquad \mathrm{Macro}_{m,A} = \frac{1}{|A|} \sum_{h \in A} q_{m,h}. \tag{14}$$

For example, Sparse RAG's Finance ZH Descriptive score is the equal-weight mean of its filing-analysis Accuracy and corpus-statistics Relaxed Accuracy@10%:

$$q_{m,\text{Fin-L-D}} = \tfrac{1}{2}\left(p_{m,\text{filing}} + p_{m,\text{statistics}}\right), \qquad L \in \{\text{ZH}, \text{EN}\}. \tag{15}$$

Finance EN Prescriptive likewise averages its capital-rule-assessment judge score and filing-fact-audit Rule Accuracy equally. Single-subset settings, such as Law EN Prescriptive, simply have $q_{m,h} = p_{m,s}$.

The Finance ZH domain–language macro then gives equal weight to the Finance ZH Descriptive, Predictive, and Prescriptive setting scores; an analytical-need macro gives equal weight to every setting assigned to that need. Thus, settings are weighted equally within each reporting group, and

subsets are weighted equally within their own setting. A subset in setting h has weight $1/(|A|\,|\mathcal{S}_h|)$ in $\mathrm{Macro}_{m,A}$; subsets from different settings need not have equal weight. This hierarchy prevents a setting with more annotated queries from receiving greater weight solely because of its size, since query counts reflect dataset construction rather than the relative importance of the underlying analytical capability. All primary scores are reported on a 0–100, higher-is-better scale, but their task-specific meanings should still be interpreted alongside the underlying metrics.

F.3 Evidence quality evaluation

Gold evidence and submitted evidence are matched within each task's corpus-specific identifier namespace. The evaluator deduplicates identifiers while preserving first occurrence for ranked metrics.

F.3.1 Population-level tasks

Legal descriptive and predictive tasks, and finance corpus statistics in both languages, require the complete qualifying population, as mentioned in Section 2.3. Let G_i be the gold set of case or company identifiers and S_i the submitted evidence set. Their per-task metrics are

$$P_i^{\mathrm{set}} = \frac{|S_i \cap G_i|}{|S_i|}, \qquad R_i^{\mathrm{set}} = \frac{|S_i \cap G_i|}{|G_i|}, \qquad F_{1,i}^{\mathrm{set}} = \frac{2P_i^{\mathrm{set}} R_i^{\mathrm{set}}}{P_i^{\mathrm{set}} + R_i^{\mathrm{set}}}. \tag{16}$$

If the gold set G_i is empty, Precision, Recall, and F1 are all set to one only when the submitted set is also empty; otherwise they are zero. If G_i is non-empty but the submitted set is empty, all three scores are zero.

F.3.2 Other tasks

For tasks requiring a compact ranked support list, let $(s_{i1}, s_{i2}, \ldots)$ be the deduplicated returned identifiers, G_i the gold set, and $h_{ir} = \mathbb{1}[s_{ir} \in G_i]$. The scorer uses

$$\begin{aligned} \mathrm{Recall@5}_i &= \frac{\sum_{r=1}^{5} h_{ir}}{|G_i|}, \qquad \mathrm{Precision@5}_i = \frac{1}{5}\sum_{r=1}^{5} h_{ir}, \\ \mathrm{RPrec}_i &= \frac{1}{|G_i|}\sum_{r=1}^{|G_i|} h_{ir}, \qquad \mathrm{MRR}_i = \begin{cases} 1/\min\{r : h_{ir} = 1\}, & \exists r : h_{ir} = 1, \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{17}$$

All ranked scores are zero when G_i is empty.

Recall@5 (R@5) and Precision@5 (P@5) are reported by default under a fixed top-5 retrieval budget. Recall@5 measures evidence coverage, while Precision@5 measures the relevance of the retrieved evidence; the cut-off balances evidence coverage against the context and processing cost of downstream analysis.

Meanwhile, two task structures motivate alternative ranked metrics. Law-EN-R uses MRR because each query has a single gold evidence item—the matched precedent—so the rank of the first relevant result directly captures retrieval quality. Legal ZH R and finance R use R-Precision because the number of gold provisions varies across queries and is usually smaller than five; evaluating the top $|G_i|$ results therefore aligns the cut-off with the amount of evidence actually required.

F.4 Efficiency evaluation

For task i, Token Usage is the total input and output tokens across all model calls apart from judge calls, including evidence supplied to a model:

$$T_i = \sum_{c \in \mathcal{C}_i} \left(T_{ic}^{\mathrm{input}} + T_{ic}^{\mathrm{output}} \right), \qquad \overline{T} = \frac{1}{N_T} \sum_{i \in I_T} T_i. \tag{18}$$

For the cost–quality analysis in Section 3.2, let M_{sb} be methods with both a reported mean token count t_m and primary answer score a_m for setting s and backbone b. Spearman's statistic is Pearson

correlation of average ranks (including tied ranks):

$$\rho_{sb} = \frac{\sum_{m \in M_{sb}} (R_t(m) - \overline{R}_t)(R_a(m) - \overline{R}_a)}{\sqrt{\sum_{m \in M_{sb}} (R_t(m) - \overline{R}_t)^2 \sum_{m \in M_{sb}} (R_a(m) - \overline{R}_a)^2}}. \tag{19}$$

A method m is dominated in M_{sb} when another method j has $t_j \leq t_m$ and $a_j \geq a_m$, with at least one strict inequality. Writing $\mathbb{1}_{\text{frontier}}(m, s, b)$ for the indicator that no such j exists, its frontier count and eligible-setting frequency are

$$C_{m,b} = \sum_{s : m \in M_{sb}} \mathbb{1}_{\text{frontier}}(m, s, b), \qquad F_{m,b} = \frac{C_{m,b}}{|\{s : m \in M_{sb}\}|}. \tag{20}$$

G Baseline Implementations and Reproduction Settings

G.1 Shared interface and models

As introduced in Section 3.1, we compare Vanilla LLM, two RAG baselines, and five adapted agents on the same 2,235 public tasks and 14 settings. Each system receives the task ID, query, instruction and permitted corpus route. It returns a final answer and ordered evidence identifiers (for population-level tasks, returns the complete selected case/company set). Gold answers and evidence remain on the scorer side.

All methods use the same API service provider for the two backbone deployments, so comparisons under a given backbone do not mix providers or endpoints. We also hold the generation parameters fixed across Vanilla LLM, both RAG baselines, and all five agents (Table 18).

Table 18: Unified model-call parameters used by all eight baselines.

Parameter	Qwen3.5-122B-A10B	DeepSeek-V4.1-Flash
`thinking`	enabled	enabled
`reasoning_effort`	–	High
`temperature`	0	0
`top_p`	1	1
`max_tokens`	8,192	8,192
`timeout`	300 seconds	300 seconds

G.2 Vanilla, sparse RAG, and dense RAG

Vanilla LLM. The model receives only the public query and instruction. Its system prompt prohibits corpus access, tools, hidden task fields, external evidence, and invented citations. This baseline therefore isolates what the backbone can produce without retrieval using its internal knowledge alone.

Sparse RAG. Sparse RAG first converts every permitted corpus route into deterministic retrieval units and stores them in a sidecar SQLite index. Long documents are split into 1,200-character chunks with 120-character overlap; structured Chinese and English financial rows are rendered as labeled fact sheets and indexed beside narrative report sections. Each unit preserves its source, parent document, retrieval ID, and benchmark evidence ID. SQLite FTS5 supplies the lexical index: English terms are lowercased and tokenized into alphanumeric words, while Chinese additionally uses overlapping character bigrams. At inference time, the public query and instruction are combined into one OR-style FTS query, and BM25 retrieves 50 units for population-level tasks (i.e. legal descriptive and financial corpus-statistics tasks) and 10 units otherwise. The highest-ranked prefix fitting an estimated 32,768-token context budget is passed to one generation call, while the full deduplicated retrieved list is retained for evidence evaluation. Sparse RAG performs no query rewriting, reranking, SQL aggregation, or iterative retrieval.

Dense RAG. Dense RAG deliberately reuses the completed Sparse RAG document stores, so its corpus routes, chunk boundaries, structured financial fact sheets, evidence IDs, top-k policy, context budget, and prompt are identical. The only changed component is retrieval. Qwen3-Embedding-4B (Zhang et al., 2025) encodes every stored unit offline and each query at inference time; embeddings are truncated to 1,024 dimensions, limited to 1,024 input tokens, and L_2 normalized. A FAISS `IndexFlatIP` then performs exact inner-product search, which is equivalent to cosine ranking for the normalized vectors. Index metadata records the route, embedding model, dimension, document count, and completion state, and the runner verifies these against the Sparse RAG document store before retrieval. Dense RAG likewise makes one retrieval pass and one answer-generation call, without query rewriting, reranking, iterative search, or structured cohort enumeration.

G.3 FIVE ADAPTED AGENTS

The five agent baselines are tied to the versions in Table 19. In every case, open-web access is replaced with task-routed local corpus tools, while the agent-specific control flow is retained.

Table 19: Versions used as the starting points for the five agent baselines.

Agent	12-character commit prefix	License
ReAct	Commit `019019be403c`	MIT
DeerFlow	Commit `e3137b13cfc2`	MIT
Tongyi DR	Commit `f72f75d8c3eb`	Apache-2.0
FinSight	Commit `8fc5bce28ac2`	GPL-3.0
LawThinker	Commit `0a9f60a05c78`	MIT

ReAct. The ReAct baseline follows a direct `search` → `open` → `report` loop. The agent runs for at most ten iterations, taking one action per iteration, with a search budget of at most five calls and at least one document inspection before answer generation. Each `search` action queries the same task-specific retrieval corpus and index used by the corresponding retrieval baseline, and `open` exposes the content of a selected returned item. Search normally returns the top 10 results, while auditing and corpus-level statistical tasks may expose up to 50 and 200 candidate identifiers, respectively. To preserve hard constraints such as company codes, reporting periods, financial metrics, and legal conditions, the original query is retained alongside the model-generated search terms. The agent may terminate early with `report` once the retrieval requirements are satisfied; otherwise, it is forced to produce the structured answer when the iteration budget is exhausted. As a fallback, at least one retrieved document is opened before answer generation, and retrieved evidence identifiers are retained even when the visible context is compacted.

DeerFlow. We adapt DeerFlow to the benchmark's closed-corpus setting using an embedded `DeerFlowClient`. Planning and subagents are disabled, and the agent is restricted to the benchmark's local retrieval tool exposed through the `web_search` interface; shell, file-system, and open-web tools are not available. Separate configurations route finance, law, and science tasks to their corresponding corpora. Search returns at most 20 results for Chinese finance and 100 for the other routes, while complete candidate identifiers are retained internally when the model-visible preview is shortened. Finance uses field-aware retrieval over financial records, legal retrieval preserves complete evidence-ID sets, and science uses full-text search over article titles and abstracts. The runner verifies the corpus route and tool configuration before inference and retries only failed or incomplete runs.

Tongyi DeepResearch. We evaluate the Tongyi DeepResearch framework rather than the separately released Tongyi-DeepResearch-30B-A3B model weights. The adaptation follows the framework's `tool_call` → `tool_response` → `answer` protocol and replaces its online research tools with a benchmark-local tool suite. This includes SQLite FTS5 with BM25 ranking over 3,000-character chunks with 200-character overlap and top-10 retrieval, paged record reading, read-only SQL, deterministic calculation, and structured population selection. The model determines the order of tool calls and uses no separate verifier. Tool outputs shown to the model are bounded, while complete SQL result sets are retained internally for evidence reporting. Deterministically recoverable formatting errors are normalized automatically, while ambiguous outputs trigger a bounded repair step.

FinSight. The FinSight adaptation retains its `DeepSearchAgent search` → `click` → `report` loop while restricting all retrieval to the benchmark's local corpus and disabling open-web access. Search returns the top 10 items for law and finance and the top 30 for science. The agent must perform both search and document inspection before reporting and runs for at most twelve iterations. Final evidence identifiers are reconstructed from the records actually accessed through the local tools, then deduplicated and normalized to the task-specific identifier format. Task-specific validators check that the final response conforms to the required answer schema, while malformed outputs undergo bounded format repair without altering the retrieved evidence trace.

LawThinker. LawThinker preserves the `Explore` → `Verify` → `Memorize` architecture. The main agent explores through the same 3,000/200-character BM25 store, read-only SQL, calculation, and population-handle tools used by Tongyi; a separate `DeepVerifier`, using the same backbone as the main agent, checks newly obtained search, document-reading, and SQL results before accepted content is stored in task-local memory. Rejected evidence is returned as corrective feedback and is never memorized. Identical accepted calls may be reused within a task, whereas changed arguments and cross-task calls require new verification. Route-specific bilingual prompts describe the relevant relational schemas and entity–attribute–value representations for each corpus. Large evidence populations are stored through local handles rather than inserted directly into the model context. If the agent reaches a final answer before the full evidence population is assembled, the answer is fixed while evidence collection is completed separately. Structured outputs are first normalized deterministically, with a repair step applied only when the format remains ambiguous.

Across the five agents, the model backbones and call parameters are controlled by Table 18; the experimental differences are therefore the control loop, available local tools, evidence-state management, validation, and stopping behavior described above. Further details on the benchmark-specific reproduction of these baselines are provided in Appendix J.

H Full Experimental Results

H.1 Complete results

Table 20 consolidates all 17 evaluation subsets into one portrait result table, with separate Qwen and DeepSeek method blocks under each setting. It retains the reported primary and secondary answer metrics, evidence metrics, and mean token usage. For Vanilla LLM, all retrieval-based metrics are unavailable because the method does not perform evidence retrieval; for population-level tasks, its outputs do not provide determinate numerical estimates without corpus access, so numerical error metrics such as MAE are also unavailable.

Table 20: All reported answer, evidence, and efficiency results, grouped by task setting and then backbone as in the main table. Bold headers identify designated primary answer metrics; bold values mark the best available score within each backbone and setting. A dash means the result is unavailable.

Method	Metrics for the indicated setting					
Law ZH descriptive						
Qwen3.5-122B-A10B						
Method	MAE ↓	**RA@10** ↑	Set P ↑	Set R ↑	Set F1 ↑	Tokens ↓
Vanilla LLM	–	0.00	–	–	–	1,818.04
Sparse RAG	80.99	6.03	20.90	32.38	15.48	24,279.56
Dense RAG	78.78	10.05	25.67	35.11	18.30	30,986.95
DeerFlow	601.65	8.00	29.03	7.07	5.76	28,587.44
ReAct	82.44	4.50	17.42	2.87	3.47	194,613.77
Tongyi DR	88.19	11.50	9.31	8.50	7.08	119,254.00
FinSight	82.66	7.50	38.70	6.28	8.20	262,575.20
LawThinker	64.77	**24.00**	44.30	57.83	38.61	224,088.61
DeepSeek-V4.1-Flash						
Method	MAE ↓	**RA@10** ↑	Set P ↑	Set R ↑	Set F1 ↑	Tokens ↓
Vanilla LLM	–	0.00	–	–	–	301.22
Sparse RAG	76.22	13.33	20.90	32.38	15.48	28,078.33
Dense RAG	79.72	11.00	26.85	37.08	19.21	28,720.46
DeerFlow	592.59	9.47	29.66	20.18	11.08	82,518.48
ReAct	76.15	10.50	36.97	12.10	13.45	924,714.28
Tongyi DR	57.18	32.00	53.43	46.92	43.78	270,857.66
FinSight	80.55	12.00	43.76	7.60	9.27	616,615.75
LawThinker	31.42	**41.00**	35.77	81.93	35.11	425,464.37
Law ZH predictive						
Qwen3.5-122B-A10B						
Method	**Rel. acc.** ↑		Set P ↑	Set R ↑	Set F1 ↑	Tokens ↓
Vanilla LLM	34.00		–	–	–	7,423.17
Sparse RAG	66.25		37.07	2.26	4.25	14,366.74
Dense RAG	62.65		58.01	3.28	6.18	12,617.38
DeerFlow	**72.15**		60.82	37.78	35.08	60,882.14
ReAct	58.23		55.87	5.43	9.33	259,936.07
Tongyi DR	70.83		6.71	3.45	3.62	178,396.18
FinSight	69.14		52.26	1.08	2.10	450,198.55
LawThinker	61.46		53.70	69.27	54.97	499,441.68
DeepSeek-V4.1-Flash						
Method	**Rel. acc.** ↑		Set P ↑	Set R ↑	Set F1 ↑	Tokens ↓
Vanilla LLM	47.00		–	–	–	1,919.85
Sparse RAG	**73.91**		37.07	2.26	4.25	13,043.34
Dense RAG	65.88		58.27	3.37	6.35	11,944.88
DeerFlow	65.06		63.51	71.79	60.91	132,820.78
ReAct	67.39		44.61	3.14	5.56	398,059.61
Tongyi DR	71.28		15.52	19.93	16.83	343,033.37
FinSight	72.63		48.62	0.74	1.46	520,889.39
LawThinker	70.71		67.87	89.90	76.11	917,339.34

Table 20 continued

Method	Metrics for the indicated setting					
Law ZH prescriptive						
Qwen3.5-122B-A10B						
Method	**Judge** ↑	R@5 ↑	R-Precision ↑			Tokens ↓
Vanilla LLM	70.34	–	–			3,128.15
Sparse RAG	65.28	12.93	10.15			4,561.19
Dense RAG	67.41	36.07	28.65			4,620.32
DeerFlow	66.11	15.33	13.50			33,218.11
ReAct	62.13	1.23	1.23			31,475.90
Tongyi DR	69.02	19.17	15.25			20,981.38
FinSight	**71.23**	25.53	23.53			37,750.20
LawThinker	70.07	21.73	18.31			53,394.03
DeepSeek-V4.1-Flash						
Method	**Judge** ↑	R@5 ↑	R-Precision ↑			Tokens ↓
Vanilla LLM	76.61	–	–			4,063.21
Sparse RAG	62.57	12.93	10.15			4,105.92
Dense RAG	75.30	32.15	24.79			3,566.99
DeerFlow	72.63	17.75	13.67			48,604.78
ReAct	74.92	34.00	30.79			26,737.13
Tongyi DR	**76.64**	18.32	15.24			110,170.96
FinSight	73.46	38.31	31.98			56,610.79
LawThinker	72.44	25.97	19.61			56,212.97
Law EN descriptive						
Qwen3.5-122B-A10B						
Method	MAE ↓	**RA@10** ↑	Set P ↑	Set R ↑	Set F1 ↑	Tokens ↓
Vanilla LLM	–	0.00	–	–	–	2,748.04
Sparse RAG	105.05	0.00	28.73	13.80	15.41	18,116.46
Dense RAG	93.11	1.04	40.64	19.93	21.95	23,883.92
DeerFlow	1333.23	15.00	31.98	18.11	8.54	31,173.08
ReAct	103.15	0.00	22.18	6.76	7.58	206,590.15
Tongyi DR	88.36	8.50	19.16	23.56	18.74	165,431.22
FinSight	104.38	0.50	59.12	4.81	7.97	230,425.68
LawThinker	79.82	**39.50**	72.46	74.00	65.28	332,118.36
DeepSeek-V4.1-Flash						
Method	MAE ↓	**RA@10** ↑	Set P ↑	Set R ↑	Set F1 ↑	Tokens ↓
Vanilla LLM	–	0.00	–	–	–	367.51
Sparse RAG	83.14	0.00	28.73	13.80	15.41	21,162.15
Dense RAG	98.66	1.00	42.42	19.30	22.00	21,444.38
DeerFlow	499.21	21.61	39.61	48.36	27.60	69,989.00
ReAct	94.51	0.50	41.98	9.14	11.14	640,288.70
Tongyi DR	43.83	46.50	69.77	62.22	61.55	268,811.71
FinSight	102.05	1.00	61.32	3.03	5.36	382,363.97
LawThinker	28.28	**53.27**	70.21	83.79	67.15	274,963.72
Law EN predictive						
Qwen3.5-122B-A10B						
Method	**Rel. acc.** ↑	Set P ↑	Set R ↑	Set F1 ↑		Tokens ↓
Vanilla LLM	15.79	–	–	–		7,689.95
Sparse RAG	**58.95**	14.66	0.32	0.61		10,818.88
Dense RAG	0.00	64.74	1.32	2.57		7,627.95
DeerFlow	56.00	64.22	27.07	27.13		68,583.93
ReAct	55.10	35.45	2.13	3.38		238,021.45
Tongyi DR	42.31	5.36	2.44	2.19		183,928.49
FinSight	48.28	62.28	0.30	0.60		255,182.73
LawThinker	51.81	67.20	77.29	64.77		715,778.47

Table 20 continued

Method	Metrics for the indicated setting					
DeepSeek-V4.1-Flash						
Method	**Rel. acc.** ↑	Set P ↑	Set R ↑	Set F1 ↑		Tokens ↓
Vanilla LLM	53.33	–	–	–		2,706.03
Sparse RAG	59.09	14.66	0.32	0.61		4,395.26
Dense RAG	55.88	63.15	1.26	2.44		7,303.28
DeerFlow	61.90	72.21	29.76	30.94		102,876.75
ReAct	54.22	47.88	1.39	2.58		263,811.17
Tongyi DR	**86.59**	93.59	90.62	89.94		309,261.72
FinSight	52.56	61.14	0.33	0.65		351,184.92
LawThinker	83.16	93.29	98.74	95.66		558,315.13
Law EN prescriptive						
Qwen3.5-122B-A10B						
Method	**Judge** ↑	R@5 ↑	MRR ↑			Tokens ↓
Vanilla LLM	76.67	–	–			4,737.35
Sparse RAG	79.28	49.12	33.41			7,864.60
Dense RAG	86.14	61.40	48.65			6,833.33
DeerFlow	80.00	21.05	11.00			93,049.81
ReAct	85.38	33.33	29.53			34,278.25
Tongyi DR	84.79	49.12	35.00			53,562.74
FinSight	85.30	40.35	37.72			220,823.19
LawThinker	**87.30**	52.63	37.07			105,406.23
DeepSeek-V4.1-Flash						
Method	**Judge** ↑	R@5 ↑	MRR ↑			Tokens ↓
Vanilla LLM	89.68	–	–			3,813.28
Sparse RAG	90.61	49.12	33.41			4,467.79
Dense RAG	92.95	64.91	45.70			4,306.91
DeerFlow	89.39	28.07	11.90			106,313.11
ReAct	**96.61**	52.63	44.12			33,110.02
Tongyi DR	91.84	49.12	35.47			101,285.57
FinSight	92.40	47.37	38.89			190,164.51
LawThinker	95.25	50.88	38.20			138,597.02
Finance ZH descriptive						
Qwen3.5-122B-A10B						
Method	SMAPE ↓	**Accuracy** ↑	Judge ↑	R@5 ↑	P@5 ↑	Tokens ↓
Vanilla LLM	50.33	46.30	17.79	–	–	4,229.04
Sparse RAG	198.92	38.89	16.38	41.62	11.80	9,116.30
Dense RAG	42.53	50.00	30.77	70.59	19.80	10,775.69
DeerFlow	0.00	63.89	23.09	63.46	17.90	49,343.75
ReAct	0.00	57.41	22.43	78.04	32.10	27,340.78
Tongyi DR	0.05	74.07	42.58	67.72	19.80	89,912.13
FinSight	0.05	**76.85**	36.58	71.95	20.10	171,096.45
LawThinker	0.00	68.52	39.64	63.43	17.20	80,884.11
DeepSeek-V4.1-Flash						
Method	SMAPE ↓	**Accuracy** ↑	Judge ↑	R@5 ↑	P@5 ↑	Tokens ↓
Vanilla LLM	8.58	59.26	23.30	–	–	1,792.64
Sparse RAG	–	33.33	17.09	41.62	11.80	5,509.21
Dense RAG	0.74	42.59	27.92	70.93	19.50	6,624.05
DeerFlow	3.21	73.15	30.74	69.32	20.40	51,512.23
ReAct	0.00	64.81	34.64	77.58	31.70	31,534.00
Tongyi DR	0.00	73.15	52.11	69.72	22.90	115,396.89
FinSight	3.60	**84.62**	40.78	96.76	24.27	170,088.00
LawThinker	0.00	74.07	50.95	63.27	16.90	71,123.85

Table 20 continued

Method	Metrics for the indicated setting					
Finance ZH statistics						
Qwen3.5-122B-A10B						
Method	MAE ↓	**RA@10** ↑	Set P ↑	Set R ↑	Set F1 ↑	Tokens ↓
Vanilla LLM	–	0.00	–	–	–	2,395.47
Sparse RAG	7.74	4.08	8.12	29.91	10.61	29,046.49
Dense RAG	5.90	6.17	13.45	43.38	17.39	43,932.38
DeerFlow	7.39	10.00	20.92	11.45	11.75	61,804.83
ReAct	4.35	**30.00**	26.76	71.07	33.40	220,012.13
Tongyi DR	3.87	29.00	22.64	36.83	24.03	139,124.45
FinSight	7.65	5.00	17.00	3.44	5.25	971,786.46
LawThinker	5.66	20.00	23.29	42.21	25.85	155,520.30
DeepSeek-V4.1-Flash						
Method	MAE ↓	**RA@10** ↑	Set P ↑	Set R ↑	Set F1 ↑	Tokens ↓
Vanilla LLM	–	0.00	–	–	–	3,686.17
Sparse RAG	5.91	11.36	8.12	29.91	10.61	28,962.56
Dense RAG	5.98	1.00	13.17	43.17	17.18	34,219.68
DeerFlow	6.64	11.00	21.12	15.05	13.75	58,154.01
ReAct	4.08	39.00	32.59	73.22	37.87	586,275.13
Tongyi DR	4.17	31.00	39.41	58.45	40.11	252,335.58
FinSight	7.26	6.00	22.75	9.41	10.79	477,930.00
LawThinker	2.46	**45.92**	32.05	89.64	36.34	361,531.82
Finance ZH predictive						
Qwen3.5-122B-A10B						
Method	**Accuracy** ↑	Judge ↑	R@5 ↑			Tokens ↓
Vanilla LLM	3.95	17.85	–			3,754.03
Sparse RAG	23.68	14.96	28.10			8,090.63
Dense RAG	27.63	28.90	41.70			12,318.00
DeerFlow	**38.16**	32.72	28.50			56,871.96
ReAct	34.21	33.79	47.40			34,526.09
Tongyi DR	36.84	32.98	51.40			137,166.66
FinSight	31.58	37.56	24.00			137,356.70
LawThinker	35.53	30.84	42.80			108,090.77
DeepSeek-V4.1-Flash						
Method	**Accuracy** ↑	Judge ↑	R@5 ↑			Tokens ↓
Vanilla LLM	44.74	41.23	–			2,662.29
Sparse RAG	22.37	18.48	28.10			4,830.21
Dense RAG	40.79	41.23	39.50			7,629.59
DeerFlow	39.47	40.38	25.40			65,632.22
ReAct	38.16	40.42	48.00			31,431.53
Tongyi DR	38.16	41.55	42.40			168,057.17
FinSight	35.53	41.39	44.10			186,027.15
LawThinker	**48.68**	42.64	40.20			137,511.23
Finance ZH prescriptive						
Qwen3.5-122B-A10B						
Method	**Judge** ↑	R@5 ↑	R-Precision ↑			Tokens ↓
Vanilla LLM	21.98	–	–			3,204.81
Sparse RAG	27.79	14.04	11.70			5,928.73
Dense RAG	31.19	33.54	28.75			5,945.40
DeerFlow	20.97	5.28	5.60			59,573.55
ReAct	32.65	21.69	16.95			33,115.54
Tongyi DR	**33.25**	23.06	18.78			60,276.53
FinSight	30.09	17.75	16.75			124,964.71
LawThinker	29.60	22.66	17.76			44,041.27

Table 20 continued

Method	Metrics for the indicated setting					
DeepSeek-V4.1-Flash						
Method	**Judge** ↑	R@5 ↑	R-Precision ↑			Tokens ↓
Vanilla LLM	39.88	–	–			2,416.00
Sparse RAG	28.71	14.04	11.70			4,292.14
Dense RAG	39.88	31.04	27.90			3,530.97
DeerFlow	**43.24**	8.08	7.21			63,957.61
ReAct	40.97	25.42	21.60			47,922.73
Tongyi DR	42.36	20.79	16.92			163,239.74
FinSight	42.43	22.36	17.43			112,624.57
LawThinker	41.93	21.99	15.78			69,598.88
Finance EN descriptive						
Qwen3.5-122B-A10B						
Method	SMAPE ↓	**Accuracy** ↑	Judge ↑	R@5 ↑	P@5 ↑	Tokens ↓
Vanilla LLM	109.65	**67.19**	12.39	–	–	3,852.07
Sparse RAG	138.84	34.38	17.80	26.73	9.20	7,808.17
Dense RAG	93.06	37.50	26.13	54.10	21.20	9,325.98
DeerFlow	55.75	31.45	40.67	70.69	43.40	55,086.13
ReAct	101.29	53.13	38.99	72.28	44.50	31,161.99
Tongyi DR	93.03	43.75	25.69	19.51	8.70	100,380.25
FinSight	91.88	32.81	32.07	52.17	18.10	359,195.82
LawThinker	79.15	46.88	40.93	31.34	11.60	289,570.91
DeepSeek-V4.1-Flash						
Method	SMAPE ↓	**Accuracy** ↑	Judge ↑	R@5 ↑	P@5 ↑	Tokens ↓
Vanilla LLM	60.97	71.88	16.48	–	–	3,269.12
Sparse RAG	78.60	46.88	21.40	26.73	9.20	4,950.86
Dense RAG	63.37	59.38	27.44	50.90	19.80	7,123.48
DeerFlow	83.43	71.88	33.12	66.58	40.00	63,624.03
ReAct	102.91	70.31	32.97	51.16	34.60	86,087.90
Tongyi DR	73.13	**76.56**	41.84	28.18	9.90	188,317.03
FinSight	103.47	65.63	32.51	54.40	19.10	316,917.00
LawThinker	88.06	75.00	50.21	36.82	14.60	250,223.14
Finance EN statistics						
Qwen3.5-122B-A10B						
Method	MAE ↓	**RA@10** ↑	Set P ↑	Set R ↑	Set F1 ↑	Tokens ↓
Vanilla LLM	–	0.00	–	–	–	1,373.94
Sparse RAG	56.74	0.00	10.94	4.81	5.24	18,436.24
Dense RAG	50.49	1.20	16.79	11.91	10.95	27,253.12
DeerFlow	54.81	2.00	20.65	14.75	10.71	64,814.03
ReAct	50.35	**24.00**	36.03	31.10	30.03	124,448.16
Tongyi DR	49.55	23.00	23.79	37.31	24.60	110,659.66
FinSight	56.21	1.00	27.00	1.81	2.94	318,851.46
LawThinker	47.49	19.19	46.89	73.60	49.75	115,129.84
DeepSeek-V4.1-Flash						
Method	MAE ↓	**RA@10** ↑	Set P ↑	Set R ↑	Set F1 ↑	Tokens ↓
Vanilla LLM	–	0.00	–	–	–	1,189.79
Sparse RAG	55.86	1.37	10.94	4.81	5.24	18,135.59
Dense RAG	56.43	0.00	17.97	12.85	11.97	23,292.13
DeerFlow	52.05	1.00	18.68	14.51	11.70	104,668.92
ReAct	48.15	11.00	37.48	17.61	16.85	989,701.15
Tongyi DR	45.04	36.00	42.16	49.42	37.08	258,436.62
FinSight	55.50	1.00	25.53	1.75	2.82	436,527.00
LawThinker	38.12	**41.41**	71.25	66.93	57.63	204,525.15

Table 20 continued

Method	Metrics for the indicated setting					
Finance EN predictive						
Qwen3.5-122B-A10B						
Method	**Accuracy** ↑	Judge ↑	R@5 ↑			Tokens ↓
Vanilla LLM	16.67	13.86	–			4,932.46
Sparse RAG	**38.33**	3.38	8.90			8,194.16
Dense RAG	20.00	18.85	42.30			10,208.98
DeerFlow	25.00	24.89	56.30			63,121.33
ReAct	25.00	23.00	51.70			38,484.30
Tongyi DR	26.67	22.75	15.50			97,041.24
FinSight	18.33	28.57	28.30			402,589.38
LawThinker	20.00	28.52	18.10			154,107.96
DeepSeek-V4.1-Flash						
Method	**Accuracy** ↑	Judge ↑	R@5 ↑			Tokens ↓
Vanilla LLM	48.33	32.63	0.00			5,529.76
Sparse RAG	**50.00**	5.70	8.90			4,804.61
Dense RAG	40.00	19.35	35.50			7,775.01
DeerFlow	31.67	32.10	43.90			58,202.42
ReAct	35.00	31.43	43.70			40,688.38
Tongyi DR	40.00	41.85	22.40			238,529.12
FinSight	31.67	35.52	25.90			312,074.34
LawThinker	38.33	44.03	21.50			185,614.34
Finance EN prescriptive						
Qwen3.5-122B-A10B						
Method	**Judge** ↑	**Rule acc.** ↑	SMAPE ↓	R@5 ↑	R-Precision ↑	Tokens ↓
Vanilla LLM	35.42	0.00	194.49	–	–	7,536.43
Sparse RAG	51.23	0.00	192.00	20.87	21.26	9,731.06
Dense RAG	55.52	1.03	158.60	36.39	33.31	6,576.78
DeerFlow	52.15	5.00	190.61	10.45	11.11	95,969.51
ReAct	51.58	**33.00**	95.28	35.79	32.61	50,728.26
Tongyi DR	**67.67**	11.00	134.01	16.05	13.18	130,601.37
FinSight	43.56	19.00	187.31	19.65	19.40	174,383.67
LawThinker	66.23	5.10	125.73	21.45	19.19	140,257.36
DeepSeek-V4.1-Flash						
Method	**Judge** ↑	**Rule acc.** ↑	SMAPE ↓	R@5 ↑	R-Precision ↑	Tokens ↓
Vanilla LLM	50.48	0.00	176.77	–	–	8,510.23
Sparse RAG	40.63	0.00	169.53	20.87	21.26	9,514.45
Dense RAG	62.41	1.00	139.26	37.57	36.11	8,212.54
DeerFlow	61.99	8.24	163.14	18.16	16.22	70,548.42
ReAct	62.65	**69.07**	64.41	38.70	34.00	102,539.02
Tongyi DR	**85.49**	36.00	83.79	26.08	21.49	257,346.10
FinSight	75.65	30.00	117.60	29.65	28.05	199,016.06
LawThinker	82.69	29.59	103.36	28.25	24.79	354,848.12
Science EN descriptive						
Qwen3.5-122B-A10B						
Method	**Token-F1** ↑	Judge ↑	R@5 ↑	P@5 ↑		Tokens ↓
Vanilla LLM	27.09	53.25	–	–		3,869.10
Sparse RAG	24.08	62.05	3.00	8.00		6,540.20
Dense RAG	28.23	63.35	10.67	35.00		7,505.80
DeerFlow	22.86	65.70	0.75	4.00		38,582.55
ReAct	24.46	53.45	6.41	22.00		28,838.25
Tongyi DR	30.49	75.95	16.44	56.00		49,217.45
FinSight	28.10	76.15	4.15	16.00		233,866.55
LawThinker	**34.59**	61.15	16.77	48.00		90,925.40

Table 20 continued

Method	Metrics for the indicated setting				
DeepSeek-V4.1-Flash					
Method	**Token-F1** ↑	Judge ↑	R@5 ↑	P@5 ↑	Tokens ↓
Vanilla LLM	43.43	70.60	–	–	1,956.00
Sparse RAG	36.92	62.20	3.00	8.00	3,240.75
Dense RAG	**46.78**	74.25	10.20	34.00	3,670.65
DeerFlow	30.67	70.30	0.47	4.00	50,696.75
ReAct	34.98	69.60	5.53	23.00	20,954.95
Tongyi DR	40.96	81.40	17.02	54.00	155,537.75
FinSight	44.63	83.60	8.06	25.00	271,801.60
LawThinker	34.62	70.30	17.23	55.00	183,292.35
Science EN predictive					
Qwen3.5-122B-A10B					
Method	**Token-F1** ↑	Judge ↑	R@5 ↑	P@5 ↑	Tokens ↓
Vanilla LLM	55.77	71.57	–	–	4,359.17
Sparse RAG	61.28	78.59	9.11	23.17	7,339.41
Dense RAG	**66.24**	90.17	15.46	41.90	5,594.03
DeerFlow	56.66	90.57	0.83	2.22	31,782.78
ReAct	62.14	80.44	15.16	36.51	23,704.54
Tongyi DR	58.64	89.65	18.18	46.67	46,521.25
FinSight	64.15	89.29	12.86	28.89	88,735.59
LawThinker	64.94	89.02	16.79	43.49	35,659.37
DeepSeek-V4.1-Flash					
Method	**Token-F1** ↑	Judge ↑	R@5 ↑	P@5 ↑	Tokens ↓
Vanilla LLM	60.95	87.73	–	–	2,561.52
Sparse RAG	62.54	78.11	9.11	23.17	3,634.51
Dense RAG	65.24	90.16	15.41	41.90	3,652.62
DeerFlow	**66.97**	90.30	1.19	3.81	46,612.30
ReAct	63.82	87.10	16.83	40.32	18,095.19
Tongyi DR	62.08	87.71	21.31	52.70	82,358.44
FinSight	60.50	86.83	11.30	26.35	125,369.40
LawThinker	58.31	84.52	20.31	49.84	66,325.06

H.2 Expanded macro-level summaries

Table 21 extends the macro-level experiment results. Its portrait panels report primary-score averages across all 14 settings, each language, each domain, each domain–language intersection, each analytical need, and each language–need intersection. Every average follows Eq. 14: designated subsets are averaged equally within a setting, then settings are averaged equally in the named group.

These views answer different descriptive questions. Under Qwen, LawThinker has the highest available overall macro (45.13), while Tongyi DR leads the Chinese-language macro (45.50). Under DeepSeek, LawThinker leads the overall macro (58.00) and law domain (69.30), while Tongyi DR leads the English-language macro (60.62). The different winners across groupings reinforce that the macro is a reporting summary of heterogeneous primary metrics, not a universal absolute measure of quality.

Table 21: Expanded primary-score macro averages (0–100, the higher the better), arranged in portrait panels. Within each panel, backbones are separate row blocks. Each task setting receives equal weight after its constituent primary subsets are averaged equally. Bold denotes the best available value within a backbone and metric.

Method	Macro group					
Overall, language, and domain						
Qwen3.5-122B-A10B						
Method	Overall	ZH	EN	Law	Finance	Science
Vanilla LLM	28.34	25.57	30.41	32.80	19.51	41.43
Sparse RAG	36.80	35.09	38.09	45.97	25.68	42.68
Dense RAG	34.02	37.84	31.16	37.88	25.75	47.23
DeerFlow	38.80	40.39	37.60	49.54	27.73	39.76
ReAct	40.60	39.24	41.62	44.22	36.07	43.30
Tongyi DR	42.65	**45.50**	40.51	47.82	**36.83**	44.56
FinSight	38.81	41.74	36.61	46.99	28.18	46.12
LawThinker	**45.13**	44.15	**45.86**	**55.69**	33.02	**49.77**
DeepSeek-V4.1-Flash						
Method	Overall	ZH	EN	Law	Finance	Science
Vanilla LLM	42.48	39.64	44.61	44.44	37.29	52.19
Sparse RAG	40.49	37.21	42.95	49.92	27.98	49.73
Dense RAG	44.13	42.44	45.41	50.34	33.98	**56.01**
DeerFlow	46.12	45.32	46.72	53.34	38.00	48.82
ReAct	48.25	47.31	48.96	50.69	45.43	49.40
Tongyi DR	56.96	52.09	**60.62**	67.48	48.27	51.52
FinSight	46.45	46.89	46.11	50.68	40.18	52.56
LawThinker	**58.00**	**55.79**	59.66	**69.30**	**50.55**	46.47
Domain by language						
Qwen3.5-122B-A10B						
Method	Law-ZH	Law-EN	Fin-ZH	Fin-EN	Sci-EN	
Vanilla LLM	34.78	30.82	16.36	22.66	41.43	
Sparse RAG	45.85	46.08	24.32	27.04	42.68	
Dense RAG	46.70	29.06	28.97	22.54	47.23	
DeerFlow	48.75	50.33	32.02	23.43	39.76	
ReAct	41.62	46.83	36.85	**35.28**	43.30	
Tongyi DR	50.45	45.20	**40.54**	33.13	44.56	
FinSight	49.29	44.69	34.20	22.17	46.12	
LawThinker	**51.84**	**59.54**	36.46	29.57	**49.77**	
DeepSeek-V4.1-Flash						
Method	Law-ZH	Law-EN	Fin-ZH	Fin-EN	Sci-EN	
Vanilla LLM	41.20	47.67	38.08	36.50	52.19	
Sparse RAG	49.94	49.90	24.47	31.48	49.73	
Dense RAG	50.73	49.94	34.16	33.80	**56.01**	
DeerFlow	49.05	57.63	41.59	34.41	48.82	
ReAct	50.94	50.44	43.68	47.17	49.40	
Tongyi DR	59.97	74.98	44.20	**52.34**	51.52	
FinSight	52.70	48.65	41.09	39.27	52.56	
LawThinker	**61.38**	**77.23**	**50.20**	50.89	46.47	
Analytical need						
Qwen3.5-122B-A10B						
Method	D	P	R			
Vanilla LLM	16.77	25.24	46.68			
Sparse RAG	13.76	**49.70**	49.49			
Dense RAG	17.35	35.30	53.25			
DeerFlow	19.91	49.59	48.91			
ReAct	22.25	46.94	55.61			
Tongyi DR	27.08	47.06	**56.60**			
FinSight	18.79	46.30	54.48			
LawThinker	**35.08**	46.75	55.66			

Table 21 continued

Method	Macro group					
DeepSeek-V4.1-Flash						
Method	D	P	R			
Vanilla LLM	21.80	50.87	57.85			
Sparse RAG	19.34	53.58	50.55			
Dense RAG	22.05	53.56	59.96			
DeerFlow	28.05	53.01	60.09			
ReAct	27.71	51.72	**69.59**			
Tongyi DR	45.56	59.62	67.90			
FinSight	27.25	50.58	65.28			
LawThinker	**49.42**	**59.84**	66.44			
Need by language						
Qwen3.5-122B-A10B						
Method	ZH-D	ZH-P	ZH-R	EN-D	EN-P	EN-R
Vanilla LLM	11.57	18.98	46.16	20.23	29.41	47.19
Sparse RAG	13.76	44.97	46.53	13.76	**52.85**	52.45
Dense RAG	19.07	45.14	49.30	16.21	28.75	57.21
DeerFlow	22.47	**55.16**	43.54	18.20	45.89	54.29
ReAct	24.10	46.22	47.39	21.01	47.41	**63.83**
Tongyi DR	31.52	53.84	**51.13**	24.12	42.54	62.06
FinSight	24.21	50.36	50.66	15.17	43.59	58.29
LawThinker	**34.13**	48.50	49.83	**35.71**	45.58	61.48
DeepSeek-V4.1-Flash						
Method	ZH-D	ZH-P	ZH-R	EN-D	EN-P	EN-R
Vanilla LLM	14.81	45.87	58.25	26.46	54.20	57.46
Sparse RAG	17.84	48.14	45.64	20.35	57.21	55.46
Dense RAG	16.40	53.33	57.59	25.82	53.71	62.33
DeerFlow	25.77	52.27	57.94	29.57	53.51	62.25
ReAct	31.20	52.77	57.95	25.38	51.01	**81.23**
Tongyi DR	42.04	54.72	**59.50**	47.91	**62.89**	76.29
FinSight	28.66	54.08	57.94	26.31	48.24	72.61
LawThinker	**50.50**	**59.69**	57.19	**48.70**	59.93	75.69

H.3 REPEATED-RUN STABILITY

We repeat two representative evaluation subsets to test whether the agent results depend on a particular execution trajectory: Law EN Predictive and the filing-analysis subset of Finance ZH Descriptive. For each method–backbone combination, three independent executions were launched with the same corpus, prompts, budgets, evaluator, and model configuration. Table 22 reports the mean and sample standard deviation. Vanilla LLM, Sparse RAG, and Dense RAG are excluded because, with temperature 0 and a fixed single-pass execution graph, their outputs are substantially more deterministic than the agents' multi-step tool-use trajectories. All five agents also use the shared decoding parameters in Table 18; their remaining run-to-run variation can arise from adaptive tool selection, intermediate state, stopping, retries, and so on.

Table 22: Repeated-run stability on two presentative evaluation subsets. Answer and evidence scores are on a 0–100 scale. Each entry is mean ± sample SD.

(a) Law-EN-Predictive

Backbone	Agent	Relation Acc.↑	Set F1↑	Tokens↓
Qwen	DeerFlow	54.69 ± 1.86	18.43 ± 12.31	$57{,}609 \pm 15{,}521$
Qwen	ReAct	56.21 ± 1.57	3.55 ± 0.23	$229{,}476 \pm 12{,}085$
Qwen	Tongyi DR	56.34 ± 6.98	48.11 ± 5.96	$336{,}399 \pm 180{,}607$
Qwen	FinSight	47.36 ± 1.30	0.64 ± 0.05	$255{,}061 \pm 172$
Qwen	LawThinker	54.34 ± 2.95	61.04 ± 3.25	$656{,}232 \pm 78{,}630$
DeepSeek	DeerFlow	58.66 ± 4.59	21.30 ± 13.63	$79{,}980 \pm 32{,}380$
DeepSeek	ReAct	55.49 ± 1.79	2.37 ± 0.30	$231{,}388 \pm 45{,}853$
DeepSeek	Tongyi DR	80.35 ± 2.20	88.23 ± 4.88	$271{,}039 \pm 27{,}669$
DeepSeek	FinSight	54.50 ± 2.74	0.67 ± 0.02	$352{,}131 \pm 1{,}339$
DeepSeek	LawThinker	83.16 ± 1.05	93.64 ± 1.75	$498{,}739 \pm 57{,}041$

(b) Finance-ZH-Descriptive: filing analysis

Backbone	Agent	Accuracy↑	Recall@5↑	Tokens↓
Qwen	DeerFlow	63.88 ± 0.02	68.01 ± 6.43	$57{,}946 \pm 12{,}165$
Qwen	ReAct	56.88 ± 0.75	73.38 ± 6.60	$37{,}148 \pm 13{,}870$
Qwen	Tongyi DR	71.60 ± 2.83	67.60 ± 0.10	$125{,}460 \pm 31{,}029$
Qwen	FinSight	78.94 ± 2.95	67.99 ± 5.61	$206{,}548 \pm 50{,}136$
Qwen	LawThinker	68.52 ± 0.93	63.39 ± 1.49	$81{,}399 \pm 3{,}999$
DeepSeek	DeerFlow	73.30 ± 0.21	68.14 ± 1.67	$51{,}653 \pm 200$
DeepSeek	ReAct	65.07 ± 0.37	71.53 ± 8.56	$36{,}834 \pm 7{,}496$
DeepSeek	Tongyi DR	74.07 ± 0.93	67.72 ± 1.74	$121{,}612 \pm 5{,}392$
DeepSeek	FinSight	83.56 ± 1.50	94.20 ± 3.63	$223{,}606 \pm 75{,}686$
DeepSeek	LawThinker	74.07 ± 0.93	65.55 ± 2.00	$126{,}519 \pm 49{,}332$

Answer stability. Primary answer scores are stable overall. Most standard deviations are below 3 points, and the principal exception is Tongyi DR with Qwen on Law EN Predictive (6.98). The relative conclusions are also preserved across runs: DeepSeek-based Tongyi DR and LawThinker remain clearly stronger on the legal association task, while FinSight remains strongest on financial filing analysis. These results suggest that, under limited computational budgets, a single ANSER-Bench run provides an acceptable estimate of answer quality, although repeated runs remain preferable when differences between methods are small.

Evidence stability. Evidence scores are also broadly stable, but they contain more visible exceptions. DeerFlow's Law EN Set F1 has standard deviations above 12 points, while ReAct's Finance ZH Recall@5 varies by 6.60–8.56 points. Evidence acquisition is directly affected by the agent's generated search terms, selected records, stopping point, and the amount of retrieved content retained across tool calls; small trajectory changes can therefore alter the submitted evidence set without proportionally changing the final answer.

Efficiency stability. Token usage is substantially less stable than either answer or evidence quality. The clearest case is Tongyi DR with Qwen on Law-EN, whose token standard deviation reaches 180,607 despite much smaller score variation. Multi-step agents can issue different numbers of searches, document reads, SQL or calculation calls, verification steps, repairs, and retries before reaching similar answers. Consequently, one run is generally sufficient for the benchmark's principal answer-quality comparison under a constrained budget, whereas precise efficiency comparisons should be interpreted more cautiously and, where possible, averaged over repeated runs.

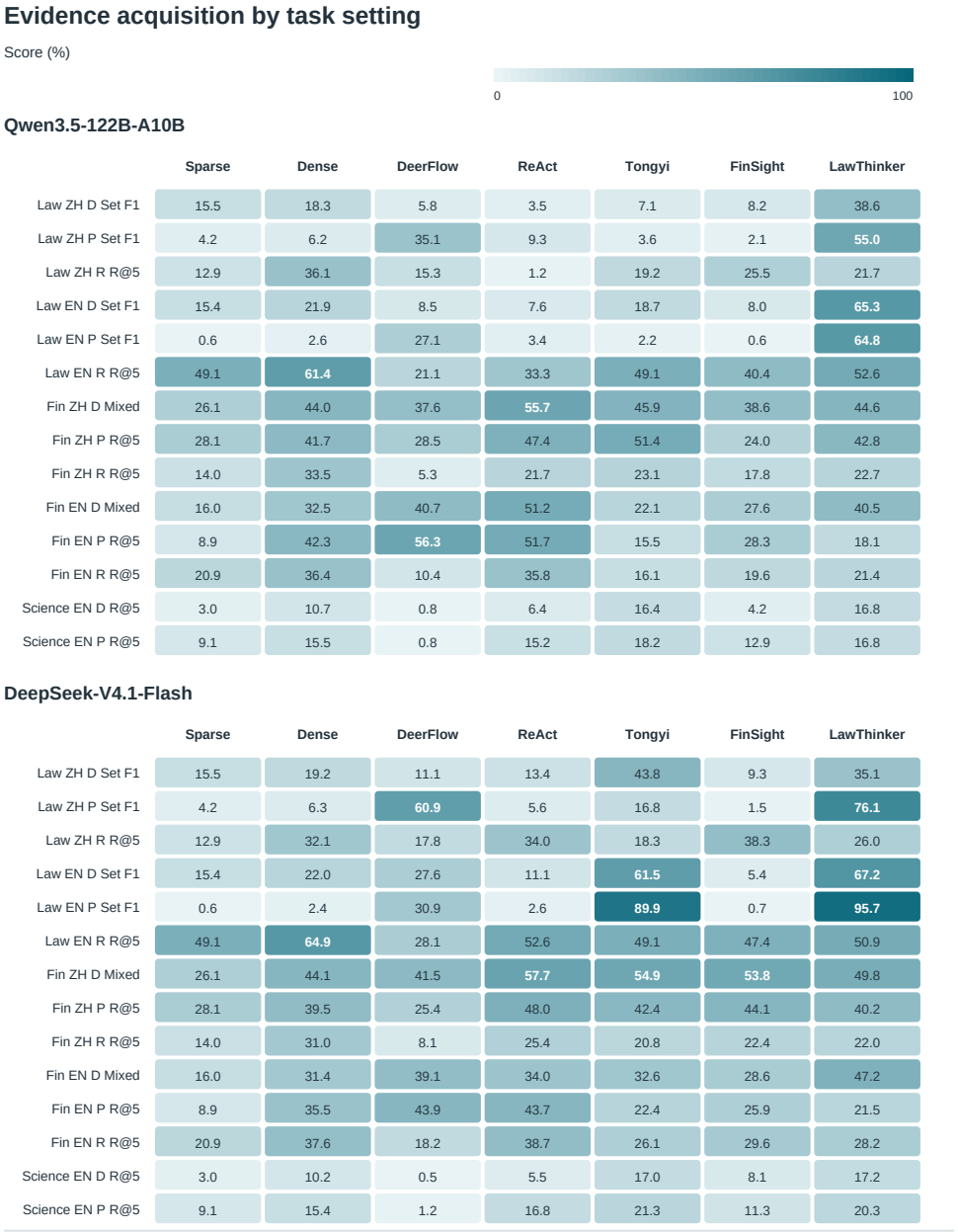


Figure 8: Evidence acquisition by task setting and backbone. Each cell is the workbook score for the designated evidence metric, on a 0–100 scale. Vanilla LLM has no retrieved evidence and is omitted. Comparisons are most direct within a row.

H.4 Evidence and efficiency profiles

Figure 8 displays the reported evidence metric across every setting and retrieval-capable method. Legal D/P rows use Set F1; finance D rows equally average filing-analysis Recall@5 and statistics Set F1; remaining rows use Recall@5. Color enables within-row method comparison. The two backbones are shown separately.

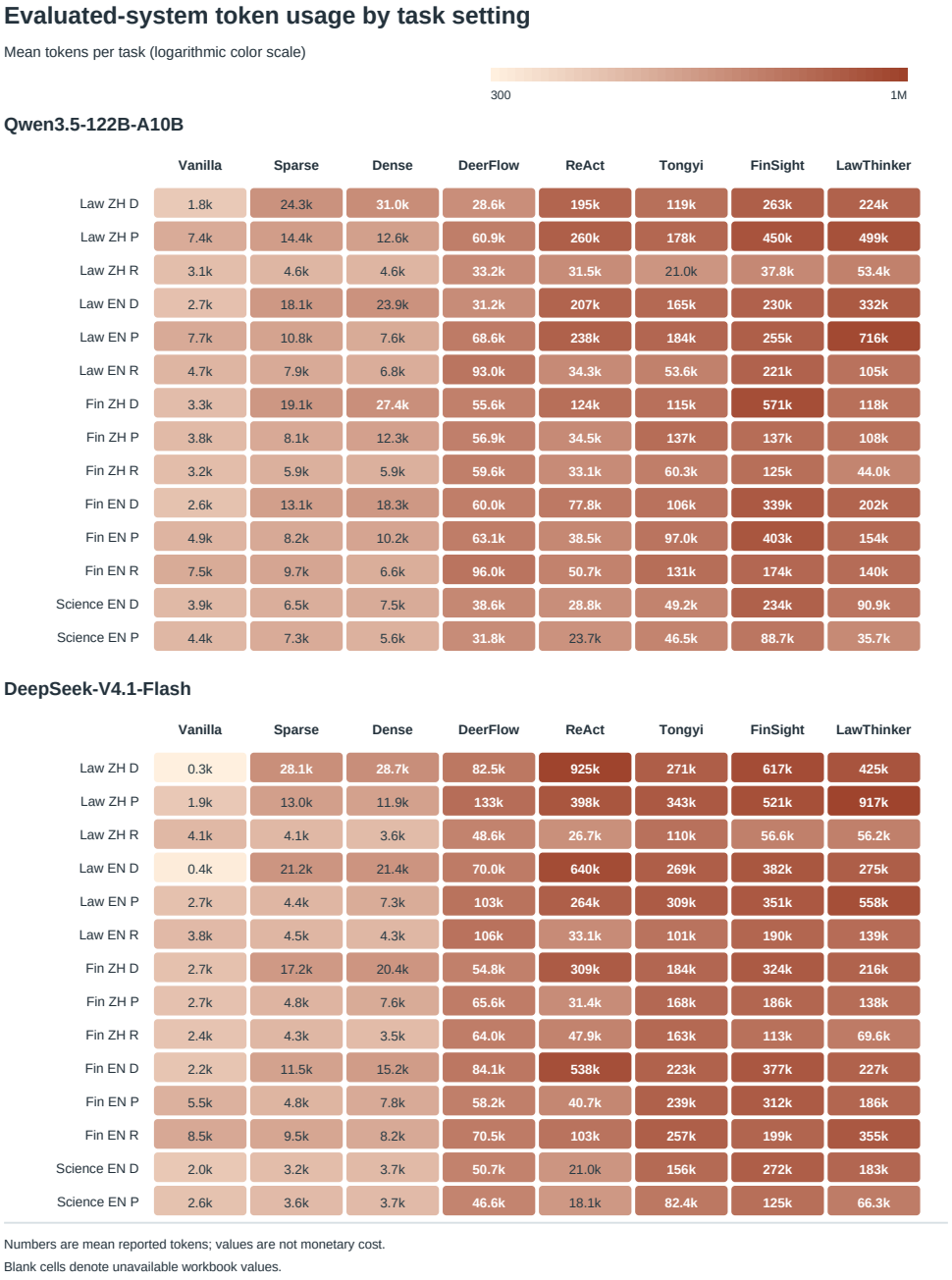


Figure 9: Mean reported token usage per task by setting, method, and backbone. Printed values use thousands of tokens; color is logarithmic.

Figure 9 displays mean token usage per task for all eight methods with a logarithmic color scale. Because all entries share the same unit, token usage can be compared not only across methods within the same row, but also across task settings, providing a directly comparable view of computational efficiency throughout the benchmark.

I Error Analysis

We inspect task-level predictions together with their submitted evidence, evaluation result, and partial execution traces. The diagnostic signals follow three stages of the analytical-search pipeline. First, *evidence-acquisition failures* concern what enters the analysis: high precision with low recall indicates cohort undercoverage, while temporal or provenance mismatch indicates retrieval from the wrong entity, source, or reporting period. Second, *evidence-use failures* arise after relevant material is available: a high evidence score paired with a low answer score is the general symptom, with aggregation or statistical errors as a common descriptive/predictive subtype and incorrect rule or precedent application as a prescriptive subtype. Third, *execution-efficiency failures* occur when repeated calls and unusually high token use do not improve either evidence coverage or the answer. These stages are related but distinct: an answer can fail because the necessary evidence was never acquired, because acquired evidence was processed incorrectly, or because a long trajectory repeatedly performs one of those unsuccessful operations. The five cases below illustrate these mechanisms.

I.1 Representative cases

Chinese queries are translated for presentation below. Evidence lists are abbreviated, and each parenthetical count gives the complete submitted set.

I.1.1 Complete cohort, incorrect statistical conclusion

Table 23: DeerFlow–DeepSeek on `law_zh_predictive_1`.

Query	In dangerous-driving cases, is causing a traffic accident associated with probation?
Evidence	（2019）鲁0103刑初518号；（2019）鲁0211刑初1195号, … (330 cases).
Evidence score	Set Precision/Recall/F1: .806/1.000/.893.
Run cost	179,462 tokens.
Answer	No clear relationship, $n = 20$, OR$= 1.000$, $p = 1.000$.
Reference	Negative relationship, $n = 266$, OR$= .526$, $p = .0146$.

Failure analysis. The run covers every normalized reference case, but the reported statistical sample contains only 20 observations. This denominator mismatch changes a significant negative relationship into "no clear relationship." Plausible causes follow from the DeerFlow design: (i) complete candidate IDs are retained internally while the model-visible search preview may be shortened, so the model may have calculated from only the visible subset; (ii) the adaptation exposes retrieval but no deterministic statistical or SQL tool, leaving population construction, odds-ratio calculation, and significance testing to the language model.

I.1.2 Precise retrieval, insufficient population coverage

Table 24: Dense RAG–DeepSeek on `law_zh_descriptive_7`.

Query	How many motor-vehicle liability cases involved personal injury reaching the level of disability?
Evidence	（2023）鲁0304民初3819号；（2019）鲁0523民初1992号, … (48 cases).
Evidence score	Set Precision/Recall/F1: .938/.050/.095.
Run cost	30,193 tokens.
Answer	28 cases.
Reference	896 cases.

Failure analysis. Most returned cases are relevant, but the ranked retrieval set covers only 48 of 896 gold cases. This pattern is consistent with three characteristics of Dense RAG: (i) population tasks use a fixed top-k of 50, which imposes a hard recall ceiling far below an 896-case population; (ii) one-shot embedding retrieval ranks semantically representative chunks but performs no structured filtering; and (iii) the single generation call may count only cases explicitly discussed in the retained context, which can explain why the answer is 28 even though 48 evidence IDs were returned.

I.1.3 EXACT EVIDENCE SET, INCORRECT COUNT AT HIGH COST

Table 25: LawThinker–DeepSeek on `law_zh_descriptive_30`.

Query	How many social-insurance disputes were won by the company because the arbitration prerequisite had not been satisfied?
Evidence	（2019）鲁1003民初603号；（2019）鲁1003民初573号, . . . (48 cases).
Evidence score	Set Precision/Recall/F1: 1.000/1.000/1.000.
Run cost	617,223 tokens.
Answer	42 cases.
Reference	48 cases.

Failure analysis. The submitted evidence exactly matches the 48-case reference set, yet the submitted answer is 42. Two LawThinker mechanisms may produce this discrepancy. First, its answer can be fixed before evidence collection finishes, after which the remaining population is assembled only for the evidence output; the answer may therefore reflect an earlier 42-case state while the final handle contains 48 cases. Second, the DeepVerifier validates newly obtained search, reading, and SQL results, but does not itself enforce a final invariant between the answer and the complete evidence handle. The 61-call Explore–Verify–Memorize trajectory also increases the chance that a stale intermediate count remains in memory after later evidence has been accepted.

I.1.4 CORRECT RETRIEVED RULE, INCORRECT APPLICATION

Table 26: ReAct–DeepSeek on `finance_en_prescriptive_232`.

Query	Audit the common-stock value for the 2023-01-31 LBSR filing (CIK `0001172178`).
Evidence	`FACT-B662...`, `FACT-78AF...`, `FACT-925B...`, and `FACT-CE27...` (4 facts); selected rule `DQC_US_0117`.
Evidence score	Recall@5/R-Precision: 1.000/1.000.
Run cost	363,692 tokens.
Answer	{`rule: 0117, value: 186`}.
Reference	{`rule: 0117, value: 1`}.

Failure analysis. The run retrieves the complete normalized evidence set and selects the same DQC rule as the reference, but maps the filing facts to an expected value of 186 rather than 1, producing SMAPE=197.86. Plausible causes are specific to the lightweight ReAct loop: (i) it provides search and document opening but no separate deterministic auditing calculator, so the model must select and transform the fact value itself; (ii) the baseline has no independent verifier after the final `report` action, so a schema-valid answer with the correct rule ID is accepted even when its numerical application is inconsistent.

I.1.5 COMPLETE FILINGS, INCORRECT TEMPORAL COMPARISON

Table 27: FinSight–Qwen on `finance_en_descriptive_87`.

Query	For Home Depot, what is the first FY2019–FY2023 filing whose Risk Factors mention supply-chain disruption: FY2019, FY2020, FY2021, or not found?
Evidence	`0000354950-24-000062`, `-22-000070`, `-20-000015`, `-21-000089`, and `-23-000059` (5 filings).
Evidence score	Precision@5/Recall@5: 1.000/1.000.
Run cost	495,959 tokens.
Answer	Not found.
Reference	FY2019.

Failure analysis. All five filings in the specified window are present, but they are returned in non-chronological accession order and the system concludes that the disclosure is absent. Three features of FinSight may contribute: (i) its search–click–report loop has no explicit fiscal-period normalization or chronological sorting stage; (ii) final evidence IDs are reconstructed from accessed records, so perfect evidence scores establish access to all five filings but do not show that their Risk Factors sections were compared consistently; and (iii) its validator checks the required answer schema rather than the semantic correspondence between an accession, its fiscal year, and the earliest qualifying disclosure. The agent can therefore return a valid option after complete retrieval while still performing the temporal comparison incorrectly.

I.2 SUMMARY

These case studies reinforce the aggregate finding that retrieval quality and answer quality are related but distinct: broader or more accurate evidence acquisition is necessary for many analytical tasks, yet it does not by itself guarantee a correct final answer. The inspected runs also provide initial hypotheses about the mechanisms behind this gap, including fixed-depth retrieval ceilings, separation between submitted evidence and the model's active analytical state, stale intermediate results, missing deterministic calculation or verification, and incomplete temporal normalization. ANSER-Bench's end-to-end design already makes this divergence, together with several plausible underlying mechanisms, visible by jointly evaluating answer quality, evidence quality, and efficiency across domains. The cases further indicate a useful direction for future research: finer-grained trajectory evaluation could identify the first erroneous step and separately assess retrieval actions, opened context, cohort-state updates, tool computations, rule application, and stopping decisions. Such extensions would complement the current benchmark and enable more precise diagnosis of how analytical-search systems succeed or fail.

J RELEASE AND REPRODUCIBILITY

J.1 RELEASE PLAN

During review, our anonymous repository already provides the complete benchmark queries together with their reference answers and evidence annotations, the evaluation code and judge prompts, and small-scale samples of the underlying corpora for inspecting the data format and retrieval interface. These materials are intended to make the benchmark construction, task schemas, and evaluation procedure directly inspectable before the full release.

Upon publication, we plan to release the complete ANSER-Bench package, including the baseline implementations and configurations used in this paper and the final corpus where redistribution is permitted. For sources that cannot be redistributed directly, we will provide stable source identifiers and reconstruction scripts. A quick-start guide, dependency specification, and commands for repro-

ducing the reported evaluation tables will accompany the release. We also plan to launch an online evaluation platform and leaderboard for standardized submission and comparison of future systems.

J.2 Versioning and documentation

Each release will carry a version number and changelog, stable task and evidence identifiers, task-to-corpus routing information, and checksums for released benchmark and corpus files. The documentation will record data provenance and applicable licenses, corpus and benchmark schemas, model and API versions, inference and evaluation settings, and the random seeds used where relevant. Reported results will be tied to a specific benchmark version so that later corrections remain distinguishable from the evaluated snapshot. The online evaluation platform will likewise record the benchmark version associated with each leaderboard submission.

J.3 Licensing and intended use

Code, original annotations, and third-party resources will be released under their respective licenses and source terms. Public legal and financial documents may still contain personal or commercially sensitive information; the release documentation will therefore state the known limitations and provide a channel for reporting problematic records or benchmark errors. ANSER-Bench is intended for research on analytical search and system evaluation, rather than for legal, financial, or clinical decision making.